IJCSIS Vol. 9 No. 10, October 2011
ISSN 1947-5500

International Journal of Computer Science & Information Security

Editorial
Message from Managing Editor

*The Journal of Computer Science and Information Security (**IJCSIS**) offers a track of quality Research & Development updates from key experts and provides an opportunity in bringing in new techniques and horizons that will contribute to advancements in Computer Science in the next few years. **IJCSIS** scholarly journal promotes and publishes original high quality research dealing with theoretical and scientific aspects in all disciplines of Computing and Information Security. Papers that can provide both theoretical analysis, along with carefully designed computational experiments, are particularly welcome.*

***IJCSIS** editorial board consists of several internationally recognized experts and guest editors. Wide circulation is assured because libraries and individuals, worldwide, subscribe and reference to **IJCSIS**. The Journal has grown rapidly to its currently level of over thousands articles published and indexed; with distribution to librarians, universities, research centers, researchers in computing, and computer scientists. After a very careful reviewing process, the editorial committee accepts outstanding papers, among many highly qualified submissions (Acceptance rate below 30%). All submitted papers are peer reviewed and accepted papers are published in the IJCSIS proceeding **(ISSN 1947-5500**). Both academia and industries are invited to submit their papers dealing with state-of-art research and future developments. IJCSIS promotes fundamental and applied research continuing advanced academic education and transfers knowledge between involved both sides of and the application of Information Technology and Computer Science.*

The journal covers the frontier issues in the engineering and the computer science and their applications in business, industry and other subjects. (See monthly Call for Papers)

Since 2009, ***IJCSIS** is published using an open access publication model, meaning that all interested readers will be able to freely access the journal online without the need for a subscription. On behalf of the editorial committee, I would like to express my sincere thanks to all authors and reviewers for their great contribution.*

Available at http://sites.google.com/site/ijcsis/

IJCSIS Vol. 9, No. 10, October 2011 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

IJCSIS EDITORIAL BOARD

TABLE OF CONTENTS

PAPR Performance analysis of DFT-spread OFDM for LTE Uplink transmission

Bader Hamad Alhasson
Department of Electrical and Computer Engineering
University of Denver
Denver, USA

Mohammad A. Matin
Department of Electrical and Computer Engineering
University of Denver
Denver, USA

Abstract— 3rd Generation Partnership Project (3GPP) LTE has adopted SC-FDMA as the uplink multiple access scheme which use single carrier modulation and frequency domain equalization. In this paper, we show that the PAPR performance of DFT-spreading technique with IFDMA can be significantly improved by varying the roll-off factor from 0 to 1 of the RC (Raised-Cosine) filter for pulse shaping after IFFT. Our PAPR reduction is 30% of DFT with IFDMA utilizing QPSK and varying the roll-off factor. We show pulse shaping does not affect LFDMA as much as it affects IFDMA. Therefore, IFDMA has an important trade-off relationship between excess bandwidth and PAPR performance since excess bandwidth increases as the roll-off factor increases. Our simulation indicates that the performance of PAPR of DFT spreading technique is dependent on the number of subcarriers assigned to each user. The effect of PAPR dependency on the method used to assign the subcarriers to each terminal is also simulated.

Index terms— Long-term-evolution (LTE); Discrete Fourier Transform (DFT); Orthogonal frequency division multiplexing (OFDM);Localized-frequency-division-multiple-access (LFDMA);Interleaved-frequency-division-multiple-access (IFDMA); peak-to-average power ratio (PAPR); single carrier frequency division multiple access (SC-FDMA).

I. INTRODUCTION

Wireless communication has experienced an incredible growth in the last decade. Two decades ago the number of mobile subscribers was less than 1% of the world's population [1]. In 2001, the number of mobile subscribers was 16% of the world's population [1]. By the end of 2001 the number of countries worldwide having a mobile network has tremendously increased from just 3% to over 90% [2]. In reality the number of mobile subscribers worldwide exceeded the number of fixed-line subscribers in 2002 [2]. As of 2010 the number of mobile subscribers was around 73% of the world's population which is around to 5 billion mobile subscribers [1].

In addition to mobile phones WLAN has experienced a rapid growth during the last decade. IEEE 802.11 a/b/g/n is a set of standards that specify the physical and data link layers in ad-hoc mode or access point for current wide use. In 1997 WLAN standard – IEEE 802.11, also known as Wi-Fi, was first developed with speeds of up to 2 Mbps [2]. At present, WLANs are capable of offering speeds up-to 600 Mbps for the IEEE 802.11n utilizing OFDM as a modulation technique in the 2.4 GHz and 5 GHz license-free industrial, scientific and medical (ISM) bands. It is important to note that WLANs do not offer the type of mobility, which mobile systems offer.
In our previous work, we analyzed a low complexity clipping and filtering scheme to reduce both the PAPR and the out-of-band-radiation caused by the clipping distortion in downlink systems utilizing OFDM technique [3]. We also modeled a mix of low mobility 1.8mph, and high mobility, 75mph with a delay spread that is constantly slighter than the guard time of the OFDM symbol to predict complex channel gains by the user by means of reserved pilot subcarriers [4]. SC-FDMA is the modified version of OFDMA. SC-FDMA is a customized form of OFDM with comparable throughput performance and complexity. The only dissimilarity between OFDM and SC-FDMA transmitter is the DFT mapper. The transmitter collects the modulation symbols into a block of N symbols after mapping data bits into modulation symbols. DFT transforms these symbols in the time domain into frequency domain. The frequency domain samples are then mapped to a subset of M subcarriers where M is greater than N. Like OFDM, an M point IFFT is used to generate the time-domain samples of these subcarriers.

OFDM is a broadband multicarrier modulation scheme where single carrier frequency division multiple access (SC-FDMA) is a single carrier modulation scheme.
Research on multi-carrier transmission started to be an interesting research area [5-7]. OFDM modulation scheme leads to better performance than a single carrier scheme over wireless channels since OFDM uses a large number of orthogonal, narrowband sub-carrier that are transmitted simultaneously in parallel ;however; high PAPR becomes an issue that limits the uplink performance more than the downlink due to the low power processing terminals. SC-FDMA adds additional advantage of low PAPR compared to OFDM making it appropriate for uplink transmission.

We investigated the channel capacity and bit error rate of MIMO-OFDM [8]. The use of OFDM scheme is the solution to the increase demand for future bandwidth-hungry wireless applications [9]. Some of the wireless technologies using OFDM are Long-Term Evolution (LTE). LTE is the standard

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

for 4G cellular technology, ARIB MMAC in Japan have adopted the OFDM transmission technology as a physical layer for future broadband WLAN systems, ETSI BRAN in Europe and Wireless local-area networks (LANs) such as Wi-Fi. Due to the robustness of OFDM systems against multipath fading, the integration of OFDM technology and radio over fiber (RoF) technology made it possible to transform the high speed RF signal to the optical signal utilizing the optical fibers with broad bandwidth [10]. Nevertheless, OFDM suffers from high peak to average power ratio (PAPR) in both the uplink and downlink which results in making the OFDM signal a complex signal [11].

The outcome of high PAPR on the transmitted OFDM symbols results in two disadvantages high bit error rate and inference between adjacent channels. This would imply the need for linear amplification. The consequence of linear amplification is more power consumption. This has been an obstacle that limits the optimal use of OFDM as a modulation and demodulation technique [12-15]. The problem of PARP affects the uplink and downlink channels differently. On the downlink, it's simple to overcome this problem by the use of power amplifiers and distinguished PAPR reduction methods. These reduction methods can't be applied to the uplink due to their difficulty in low processing power devices such as mobile devices. On the uplink, it is important to reduce the cost of power amplifiers as well.

PAPR reduction schemes have been studied for years [16-19]. Some of the PAPR reduction techniques are: Coding techniques which can reduce PAPR at the expense of bandwidth efficiency and increase in complexity [20-21]. The probabilistic technique which includes SLM, PTS, TR and TI can also reduce PAPR; however; suffers from complexity and spectral efficiency for large number of subcarriers [22-23].

We show the effect of PAPR dependency on the method used to assign the subcarriers to each terminal. PAPR performance of DFT-spreading technique varies depending on the subcarrier allocation method.

II SYSTEM CONFIGURATION OF SC-FDMA and OFDMA

SC-FDMA:

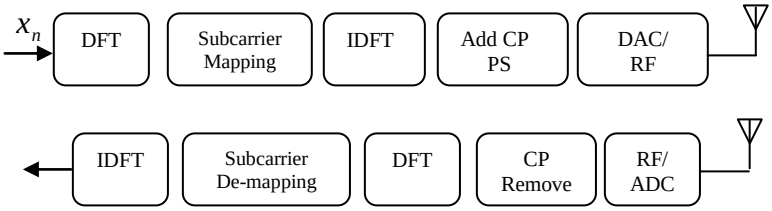


Fig.1. Transmitter and receiver structure of SC-FDMA

The transmitters in Figure 1 and 2 perform some signal-processing operations prior to transmission. Some of these operations are the insertion of cyclic prefix (CP), pulse shaping (PS), mapping and the DFT. The transmitter in Figure 1 converts the binary input signal to complex subcarriers. In a SC-FDMA, DFT is used as the first stage to modulate subcarriers. DFT produce a frequency domain representation of the input signal.

OFDMA:

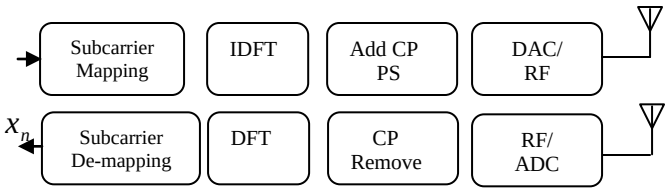


Fig.2. Transmitter and receiver structure of OFDMA

Figure 2 illustrates the configuration of OFDMA transmitter and receiver. The only difference between SC-FDMA and OFDMA is the presences of the DFT and IDFT in the transmitter and receiver respectively of SC-FDMA. Hence, SC-FDMA is usually referred to as DFT-spread OFDMA.

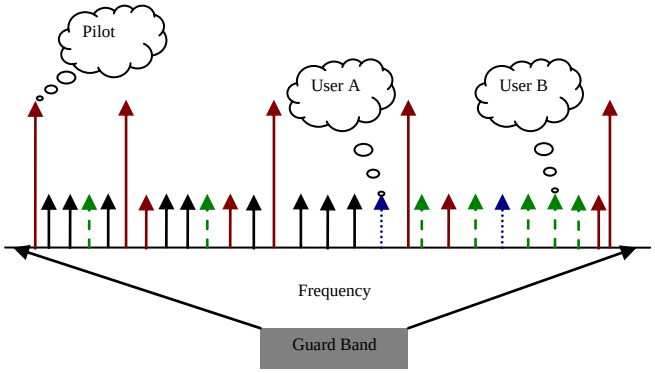


Fig. 1. OFDM available bandwidth is divided into subcarriers that are mathematically orthogonal to each other [3]

II. SYSTEM MODEL

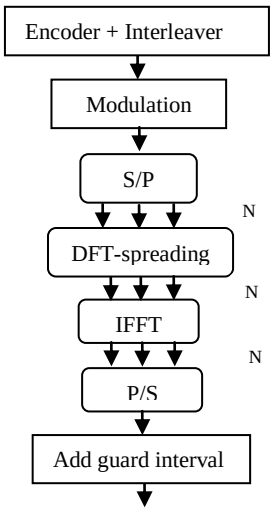


Fig. 2.DFT-spreading OFDM single carrier transmitter

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

One of the major drawbacks of OFDM is the high peak-to-average power ratio (PAPR) of the transmitted signals, i.e., the large variations in the instantaneous power of the transmitted signal. This would require linear amplification. The result of such linear amplification would imply more power consumption. This is significant on the uplink, due to the low mobile-terminal power consumption and cost. Therefore, wide-band single-carrier transmission is an alternative to multi-carrier transmission, particularly for the uplink. One of such single-carrier transmission scheme can be implemented using DFT-spread OFDM which has been selected as the uplink transmission scheme for LTE allowing for small variations in the instantaneous power of the transmitted uplink signal.

The main advantage of DFTS-OFDM, compared to OFDM, is the reduction of variations in the instantaneous transmit power, leading to the possibility for increased power-amplifier efficiency.

DFT spreading technique is a promising solution to reduce PAPR because of it's superiority in PAPR reduction performance compared to block coding, Selective Mapping (SLM), Partial Transmit Sequence (PTS) and Tone Reservation (TR) [24-25]. SC-FDMA and OFDMA are both multiple-access versions of OFDM. There are two subcarrier mapping schemes in single carrier frequency division multiple access (SC-FDMA) to allocate subcarriers between units: Distributed FDMA and Localized FDMA.

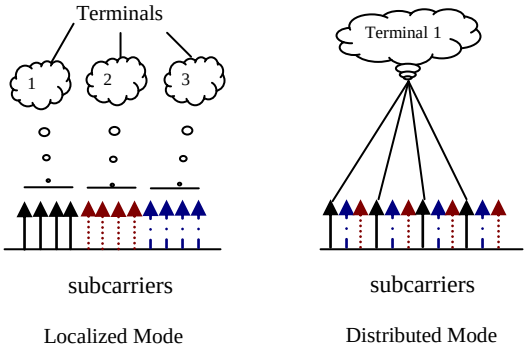


Fig. 3.Subcarrier allocation methods for multiple users (3 users, 12 subcarriers, and 4 subcarriers allocated per user).

III SIMULATION AND RESULTS

Before examining the reduction of PAPR, let us consider a single-carrier system where N=1. Figure 4 shows both the baseband QPSK-modulated signal and the passband signal with a single carrier frequency of 1 Hz and an oversampling factor of 8. Figure 4a shows that the baseband signal's average and peak power values are the same that is PAPR is 0dB.

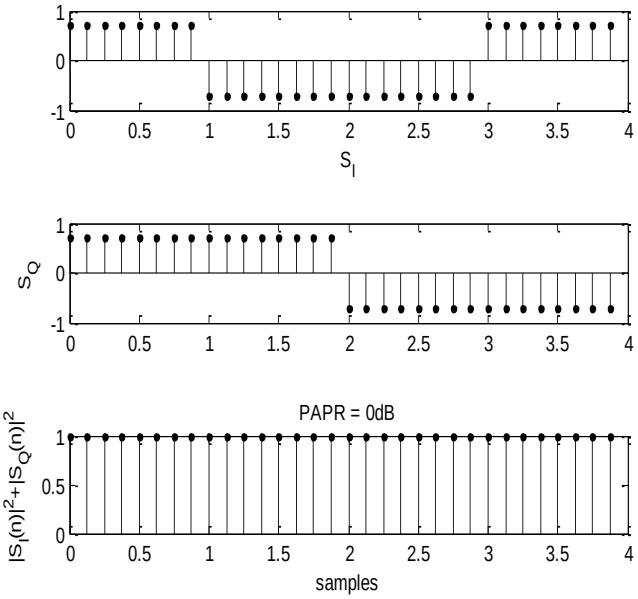


Fig. 4. (a) Baseband signal

On the other hand, Figure 4b shows the passband signal with a PAPR of 3,01 dB.

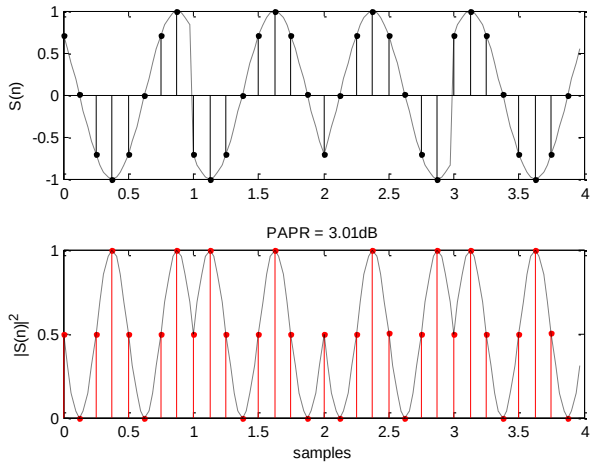


Fig. 4. (b) Passband signal

Note that the PAPR varies in the passband signal depending on the carrier frequency. As a result, when measuring the PAPR of a single-carrier system, then we must be taken into consideration the carrier frequency of the passband signal.

A. *Interleaved, Localized and Orthogonal-FDMA*

There are two channel allocation schemes for SC-FDMA systems; i.e., the localized and interleaved schemes where the subcarriers are transmitted subsequently, rather than in parallel. In the following simulation results, we compared

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

different allocation schemes of SC-FDMA systems and their PAPR. These types of allocation schemes are subject to intersymbol interference when the signal suffers from sever multipath propagation. In SC-FDMA this type of interference can be substantial and usually an adaptive frequency domain equalizer is placed at the base station. This type of arrangement makes sense in the uplink of cellular systems due to the additional benefit that SC-FDMA adds in terms of PAPR. In this type of arrangement, i.e, single carrier system the burden of linear amplification in portable terminals is shifted to the base station at the cost of complex signal processing, that is frequency domain equalization.

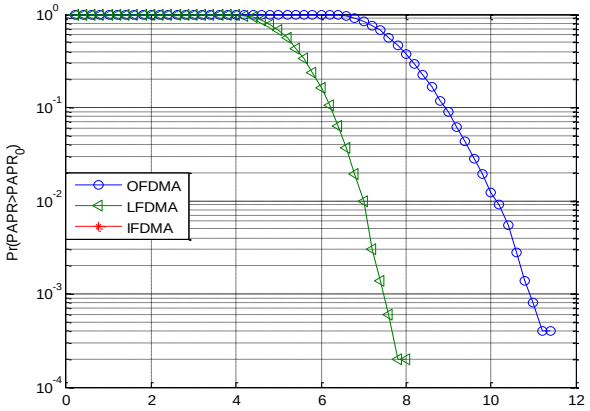


Fig. 4. (a) QPSK

Figure 4 show the performance of PAPR while the number of subcarriers is 256 and the number of subcarriers assigned to each unit or mobile device is 64. This simulation helps in evaluating the performance of PAPR with different mapping schemes and modulation techniques. In LFDMA each user transmission is localized in the frequency domain where in the DFDMA each user transmission is spread over the entire frequency band making it less sensitive to frequency errors and diversifies frequency.

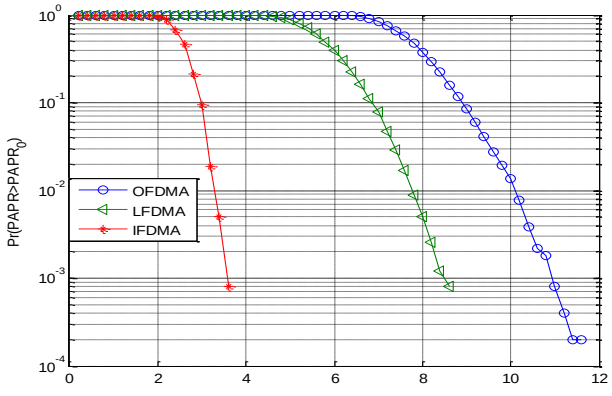


Fig. 4. (b) 16 QAM

The three figures of 4 show that when the single carrier is mapped either by LFDMA or DFDMA, it outperforms OFDMA due to the fact that in an uplink transmission, mobile terminals work differently then a base station in terms of power amplification. In the uplink transmission PAPR is more of a significant problem then on the downlink due to the type and capability of the amplifiers used in base station and mobile devices. For instance, when a mobile circuit's amplifier operates in the non-linear region due to PAPR, the mobile devise would consume more power and become less power efficient whereas base stations don't suffer from this consequence. Therefore, OFDM works better in the downlink transmission in terms of PAPR.

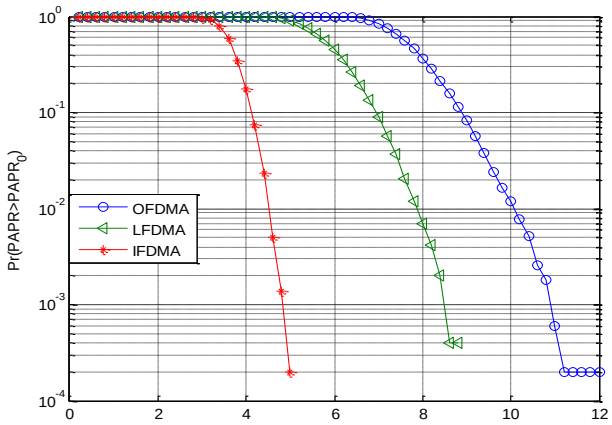


Fig. 4. (c) 64 QAM

Our results show the effect of using Discrete Fourier Transform spreading technique to reduce PAPR for OFDMA, LFDMA and OFDMA with N=256 and N_{unit} =64. A comparison is shown in Figure 4 a,b and c utilizing different modulation schemes. The reduction in PAPR is significant when DFT is used. For example, Figure 4(b) where Orthogonal-FDMA, Localized-FDMA and Interleaved-FDMA have the values of 3.9 dB, 8.5 dB and 11 dB, respectively. The reduction of PAPR in IFDMA utilizing the DFT-spreading technique compared to OFDMA without the use of DFT is 6.1 dB. Such reduction is significant in the performance of PAPR. Based on the simulation results in Figure 2 we can see that single carrier frequency division multiple access systems with Interleaved-FDMA and Localized-FDMA perform better than OFDMA in the uplink transmission. Although Interleaved-FDMA performs better than OFDMA and LFDMA, LFDMA is preferred due to the fact that assigning subcarriers over the whole band of IFDMA is complicated while LFDMA doesn't require the insertion of pilots of guard bands.

B. *Pulse shaping*

The idea of pulse shaping is to find an efficient transmitter and a corresponding receiver waveform for the current channel

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

condition [26]. The raised-cosine filter is used for pulse shaping because it is able to minimize intersymbol interference (ISI). In this section we show the effect of pulse shaping on the PAPR. Figure 4 a and b show the PAPR performance of both IFDMA and LFDMA, varying the roll-off-factor of the raised cosine filter for pulse shaping after IFFT. The roll-off-factor is a measure of excess bandwidth of the filter. The raised cosine filter can be expressed as:

$$p(t) = \frac{\sin(\pi t/T)}{\pi t/T} \cdot \frac{\cos(\pi \alpha t/T)}{1-4\alpha^2 t^2/T^2}$$

Where T is the symbol period and α is the roll-off factor.

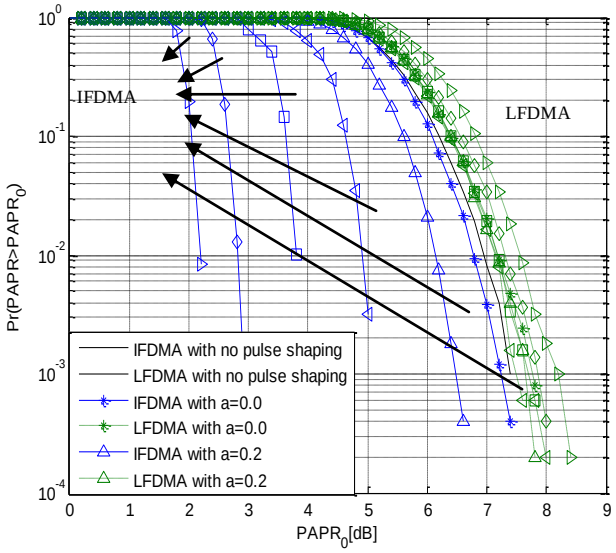


Fig. 5. (a) QPSK

Figures 5 a and b imply that IFDMA is more sensitive to pulse shaping than LFDMA. The PAPR performance of the IFDMA is greatly improved by varying the roll-off factor from 0 to 1. On the other hand LFDMA is not affected so much by the pulse shaping.

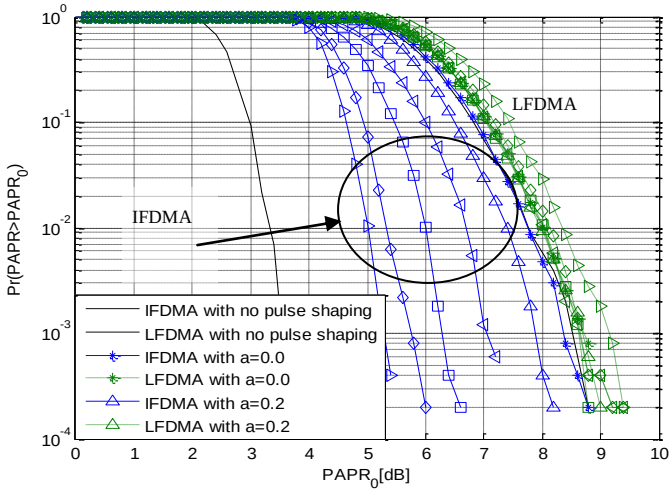


Fig. 5. (b) 16 QAM

It is important to note that IFDMA has a trade-off relationship between excess bandwidth and PAPR performance because any excess in bandwidth increases as the roll-off factor increases. Excess bandwidth of a filter is the bandwidth occupied beyond the Nyquist bandwidth.

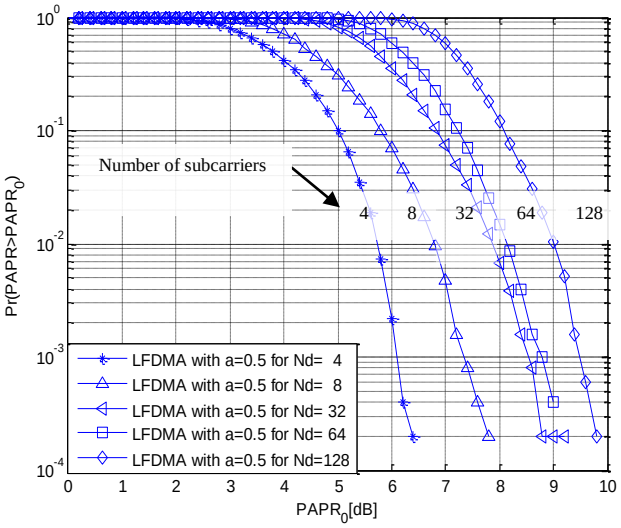


Fig. 6. PAPR performance of DFT-spreading technique when the number of subcarriers vary

The PAPR performance of the DFT-spreading technique depends on the number of subcarriers allocated to each user. Figure 5 shows the performance of DFT-spreading for LFDMA with a roll-off factor of 0.5. The degraded performance by about 3.5 dB can be seen as the number of subcarriers increase from 4 to 128 subcarriers.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

V. CONCLISION

We have shown the importance of the trade-off relationship of IFDMA between excess bandwidth and PAPR performance due to the fact that any excess in bandwidth increases as the roll-off factor increases. Our results show The PAPR performance of the IFDMA is greatly improved by varying the roll-off factor. On the other hand LFDMA is not affected so much by the pulse shaping. It was also shown that a SC-FDMA system with Interleaved-FDMA or Localized FDMA performs better than Orthogonal-FDMA in the uplink transmission where transmitter power efficiency is of great importance in the uplink. LFDMA and IFDMA result in lower average power values due to the fact that OFDM and OFDMA map their input bits straight to frequency symbols where LFDMA and IFDMA map their input bits to time symbols. We conclude that single carrier-FDMA is a better choice on the uplink transmission for cellular systems. Our conclusion is based on the better efficiency due to low PAPR and on the lower sensitivity to frequency offset since SC-FDMA has a maximum of two adjacent users. Finally yet importantly, the PAPR performance of DFT-spreading technique degrades as the number of subcarriers increase.

REFERENCES

[1] ITU, “World Telecommunication Development Report 2002: Reinventing Telecoms”, March 2002.

[2] Anthony Ng’oma, “Radio-over-Fibre Technology for Broadband Wireless Communication Systems”, June 2005.

[3] Bader Alhasson, and M. Matin “Reduction of PAPR for OFDM Downlink and IFDMA Uplink Wireless Transmissions”, International Journal of Computer Science and Information Security,Vol. 9, No. 3, March 2011

[4] Bader Alhasson, and M. Matin “The challenge of scheduling user transmissions on the downlink of a long-term evolution (LTE) cellular communication system”, Proc. SPIE, Vol. 7797, 779719, Sep 2010.

[5] H. Atarashi, S. Abeta, and M. Sawahashi, “Variable spreading factor orthogonal frequency and code division multiplexing (VSF-OFCDM) for broadband packet wireless access,” IEICE Trans. Commun., vol. E86-B, pp. 291-299, Jan. 2003.

[6] R. Kimura and F. Adachi, “Comparison of OFDM and multicode MC-CDMA in a frequency selective fading channel, ” IEE Electronics Letters, vol. 39, no.3, pp. 317-318, Feb. 2003.

[7] Z Wang and G. B. Giannakis, “Complex-field Coding for OFDM over Fading Wireless Channels,” IEEE Trans. Inform. Theory, vol. 49, pp.707-720, Mar. 2003.

[8] Alhasson Bader, Bloul A., Li X., and Matin M. A: “LTE-advanced MIMO uplink for mobile system” Proc. SPIE, Vol. 7797, 77971A, 2010.

[9] L. Mehedy, M. Bakaul, A. Nirmalathas, "115.2 Gb/s optical OFDM transmission with 4 bit/s/Hz spectral efficiency using IEEE 802.11a OFDM PHY," in proc. the 14th OptoElectronics and Communications Conference, 2009 (OECC 2009), July 2009.

[10] Alhasson Bader, Bloul A., and Matin M. A.: “Dispersion and Nonlinear Effects in OFDM-RoF system”, SPIE, Vol. 7797, 779704, 2010.

[11] J. Tellado, “Multicarrier transmission with low PAR,” Ph.D. dissertation,Stanford Univ., Stanford, CA, 1998.

[12] Z.-Q. Luo and W. Yu, “An introduction to convex optimization for communications and signal processing," IEEE J. Sel. Areas Communication, vol. 24, no. 8, pp. 1426-1438, Aug. 2006.

[13] J. Tellado, “Peak to average power reduction for multicarrier modulation," Ph.D. dissertation, Stanford University, Stanford, USA, 2000.

[14] A. Aggarwal and T. Meng, “Minimizing the peak-to-average power ratio of OFDM signals using convex optimization," IEEE Trans. Signal Process., vol. 54, no. 8, pp. 3099-3110, Aug. 2006.

[15] Y.-C. Wang and K.-C. Yi, “Convex optimization method for quasiconstant peak-to-average power ratio of OFDM signals," IEEE Signal Process. Lett., vol. 16, no. 6, pp. 509-512, June 2009.

[16] S. H. Wang and C. P. Li, “A low-complexity PAPR reduction scheme for SFBC MIMO-OFDM systems,” IEEE Signal Process. Lett., vol. 16, no. 11, pp. 941–944, Nov. 2009.

[17] J. Hou, J. Ge, D. Zhai, and J. Li, “Peak-to-average power ratio reduction of OFDM signals with nonlinear companding scheme,” IEEE Trans. Broadcast., vol. 56, no. 2, pp. 258–262, Jun. 2010.

[18] T. Jaing, W. Xiang, P. C. Richardson, D. Qu, and G. Zhu, “On the nonlinear companding transform for reduction in PAPR of MCM,” IEEE Trans. Wireless Commun., vol. 6, no. 6, pp. 2017–2021, Jun. 2007.

[19] S. H. Han and J. H. Lee, “An overview of peak-to-average power ratio reduction techniques for multicarrier transmission,” IEEE Wireless Commun., vol. 12, pp. 56–65, Apr. 2005.

[20] Wilkinson, T.A. and Jones, A.E . Minimization of the peak-to-mean envelope power ratio of multicarrier transmission scheme by block coding. IEEE VTC’95, Chicago, vol. 2,pp. 825-829. July, 1995.

[21] Park, M.H. PAPR reduction in OFDM transmission using Hadamard transform. IEEE ICC’00, vol.1, pp. 430-433. 2000

[22] Bauml, R.W., Fischer, R.F.H., and Huber, J.B. Reducing the peak to average power ratioof multicarrier modulation by selective mapping. Electron. Lett., 32(22), 2056-2057. 1996.

[23] Muller, S.H and Huber, J.B. a novel peak power reduction scheme for OFDM. PIMRC, vol. 3,pp. 1090-1094. 1997.

[24] H. G. Myung, J. Lim, and D. J. Goodman, “Single Carrier FDMA for Uplink Wireless Transmission,” IEEE Vehicular Technology Mag., vol.1, no. 3, pp. 30 – 38, Sep. 2006.

[25] H. G. Myung and David J. Goodman, Single Carrier FDMA”, WILEY, 2008.

[26] N.J. Baas and D.P. Taylor, “Pulse shaping for wireless communication over timeor frequency-selective channels", IEEE Transactions on Communications, vol 52, pp.1477-1479, Sep. 2004.

[27] Bloul A., Mohseni S., Alhasson Bader, Ayad M., and Matin M. A.: “Simulation of OFDM technique for wireless communication systems”, Proc. SPIE, Vol. 7797, 77971B, 2010.

[28] Cho, kim, Yang & Kang. MIMO-OFDM Wireless Communications with MATLAB. IEEE Press 2010.

AUTHORS PROFILE

Bader Hamad Alhasson is a PhD candidate from the University of Denver. He received a bachelor degree in Electrical Engineering (EE) in 2003 from the University of Colorado at Denver (UCD) in the United States, a Master's of Science in EE and a Master's of Business Administration (MBA) in 2007 from UCD. His primary research interest is in the optimization of OFDM as a modulation and multiplexing scheme.

Dr. Mohammad Abdul Matin, Associate Professor of Electrical and Computer Engineering, in the School of Engineering and Computer Science, University of Denver. He is a Senior Member of IEEE & SPIE and member of OSA, ASEE and Sigma Xi. His research interest is in Optoelectronic Devices (such as Sensors and Photovoltaic) RoF, URoF, Digital, Optical & Bio-Medical Signal & image Processing, OPGW, Engineering Management and Pedagogy in Engineering Education.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Effect of Curvature on the Performance of Cylindrical Microstrip Printed Antenna for TM_{01} mode Using Two Different Substrates

Ali Elrashidi
Department of Computer and Electrical Engineering
University of Bridgeport
Bridgeport, CT, USA
aelrashi@bridgeport.edu

Khaled Elleithy
Department of Computer and Electrical Engineering
University of Bridgeport
Bridgeport, CT, USA
elleithy@bridgeport.edu

Hassan Bajwa
Department of Electrical Engineering line 2: name of
University of Bridgeport
Bridgeport, CT, USA
hbjwa@bridgeport.edu

Abstract— **Curvature has a great effect on fringing field of a microstrip antenna and consequently fringing field affects effective dielectric constant and then all antenna parameters. A new mathematical model for input impedance, return loss, voltage standing wave ratio and electric and magnetic fields is introduced in this paper. These parameters are given for TM_{01} mode and using two different substrate materials RT/duroid-5880 PTFE and K-6098 Teflon/Glass. Experimental results for RT/duroid-5880 PTFE substrate are also introduced to validate the new model.**

Keywords: Fringing field, Curvature, effective dielectric constant and Return loss (S11), Voltage Standing Wave Ratio (VSWR), Transverse Magnetic TM_{01} mode.

I. INTRODUCTION

Due to the unprinted growth in wireless applications and increasing demand of low cost solutions for RF and microwave communication systems, the microstrip flat antenna, has undergone tremendous growth recently. Though the models used in analyzing microstrip structures have been widely accepted, the effect of curvature on dielectric constant and antenna performance has not been studied in detail. Low profile, low weight, low cost and its ability of conforming to curve surfaces [1], conformal microstrip structures have also witnessed enormous growth in the last few years. Applications of microstrip structures include Unmanned Aerial Vehicle (UAV), planes, rocket, radars and communication industry [2]. Some advantages of conformal antennas over the planer microstrip structure include, easy installation (randome not needed), capability of embedded structure within composite *aerodynamic* surfaces, better angular coverage and controlled gain, depending upon shape [3, 4]. While Conformal Antenna provide potential solution for many applications, it has some drawbacks due to bedding [5]. Such drawbacks include phase, impedance, and resonance frequency errors due to the stretching and compression of the dielectric material along the inner and outer surfaces of conformal surface. Changes in the dielectric constant and material thickness also affect the performance of the antenna. Analysis tools for conformal arrays are not mature and fully developed [6]. Dielectric materials suffer from cracking due to bending and that will affect the performance of the conformal microstrip antenna.

II. BACKGROUND

Conventional microstrip antenna has a metallic patch printed on a thin, grounded dielectric substrate. Although the patch can be of any shape, rectangular patches, as shown in Figure 1 [7], are preferred due to easy calculation and modeling.

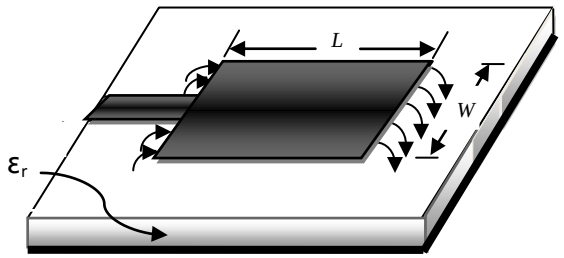


FIGURE 1. Rectangular microstrip antenna

Fringing fields have a great effect on the performance of a microstrip antenna. In microstrip antennas the electric filed in the center of the patch is zero. The radiation is due to the fringing field between the periphery of the patch and the ground plane. For the rectangular patch shown in the Figure 2, there is no field variation along the width and thickness. The amount of the fringing field is a function of the dimensions of the patch and the height of the substrate. Higher the substrate, the greater is the fringing field.

Due to the effect of fringing, a microstrip patch antenna would look electrically wider compared to its physical dimensions. As shown in Figure 2, waves travel both in

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

substrate and in the air. Thus an effective dielectric constant εreff is to be introduced. The effective dielectric constant εreff takes in account both the fringing and the wave propagation in the line.

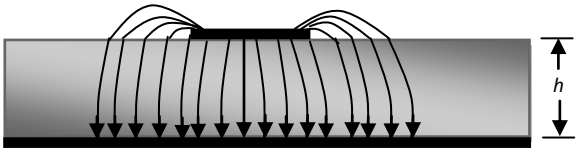


FIGURE 2. Electric field lines (Side View).

The expression for the effective dielectric constant is introduced by A. Balanis [7], as shown in Equation 1.

$$\varepsilon_{reff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left[1 + 12\frac{h}{w}\right]^{-\frac{1}{2}} \quad (1)$$

The length of the patch is extended on each end by ΔL is a function of effective dielectric constant ε_{reff} and the width to height ratio (W/h). ΔL can be calculated according to a practical approximate relation for the normalized extension of the length [8], as in Equation 2.

$$\frac{\Delta L}{h} = 0.412\frac{(\varepsilon_{reff} + 0.3)\left(\frac{W}{h} + 0.264\right)}{(\varepsilon_{reff} - 0.258)\left(\frac{W}{h} + 0.8\right)} \quad (2)$$

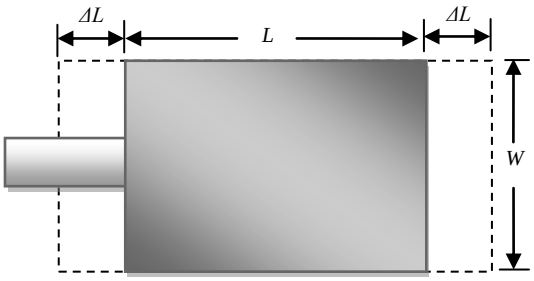


FIGURE 3. Physical and effective lengths of rectangular microstrip patch.

The effective length of the patch is L_{eff} and can be calculated as in Equation 3.

$$L_{eff} = L + 2\Delta L \quad (3)$$

By using the effective dielectric constant (Equation 1) and effective length (Equation 3), we can calculate the resonance frequency of the antenna f and all the microstrip antenna parameters.

Cylindrical-Rectangular Patch Antenna

All the previous work for a conformal rectangular microstrip antenna assumed that the curvature does not affect the effective dielectric constant and the extension on the length. The effect of curvature on the resonant frequency has been presented previously [9]. In this paper we present the effect of fringing field on the performance of a conformal patch antenna. A mathematical model that includes the effect of curvature on fringing field and on antenna performance is presented. The cylindrical-rectangular patch is the most famous and popular conformal antenna. The manufacturing of this antenna is easy with respect to spherical and conical antennas.

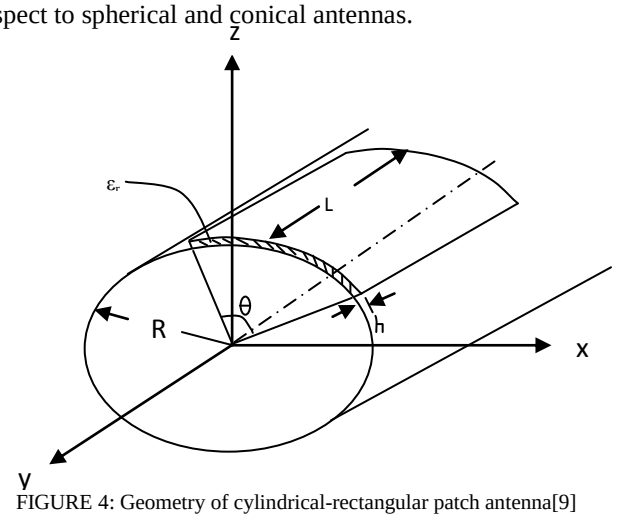


FIGURE 4: Geometry of cylindrical-rectangular patch antenna[9]

Effect of curvature of conformal antenna on resonant frequency been presented by Clifford M. Krowne [9, 10] as:

$$(f_r)_{mn} = \frac{1}{2\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{m}{2\theta a}\right)^2 + \left(\frac{n}{2b}\right)^2} \quad (4)$$

Where $2b$ is a length of the patch antenna, a is a radius of the cylinder, 2θ is the angle bounded the width of the patch, ε represents electric permittivity and μ is the magnetic permeability as shown in Figure 4.

Joseph A. *et al*, presented an approach to the analysis of microstrip antennas on cylindrical surface. In this approach, the field in terms of surface current is calculated, while considering dielectric layer around the cylindrical body. The assumption is only valid if radiation is smaller than stored energy[11]. Kwai *et al*. [12]gave a brief analysis of a thin cylindrical-rectangular microstrip patch antenna which includes resonant frequencies, radiation patterns, input impedances and Q factors. The effect of curvature on the characteristics of TM_{10} and TM_{01} modes is also presented in Kwai *et al*. paper. The authors first obtained the electric field under the curved patch using the cavity model and then calculated the far field by considering the equivalent magnetic current radiating in the presence of cylindrical surface. The cavity model, used for the analysis is only valid for a very thin dielectric. Also, for much small thickness than a wavelength and the radius of curvature, only *TM* modes are assumed to exist. In order to calculate the radiation patterns of cylindrical-rectangular patch antenna. The authors introduced the exact Green's function approach. Using Equation (4), they obtained expressions for the far

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

zone electric field components E_θ and E_φ as a functions of Hankel function of the second kind $H_p^{(2)}$. The input impedance and *Q* factors are also calculated under the same conditions.

Based on cavity model, microstrip conformal antenna on a projectile for GPS (Global Positioning System) device is designed and implemented by using perturbation theory is introduced by Sun L., Zhu J., Zhang H. and Peng X [13]. The designed antenna is emulated and analyzed by IE3D software. The emulated results showed that the antenna could provide excellent circular hemisphere beam, better wide-angle circular polarization and better impedance match peculiarity.

Nickolai Zhelev introduced a design of a small conformal microstrip GPS patch antenna [14]. A cavity model and transmission line model are used to find the initial dimensions of the antenna and then electromagnetic simulation of the antenna model using software called FEKO is applied. The antenna is experimentally tested and the author compared the result with the software results. It was founded that the resonance frequency of the conformal antenna is shifted toward higher frequencies compared to the flat one.

The effect of curvature on a fringing field and on the resonance frequency of the microstrip printed antenna is studied in [15]. Also, the effect of curvature on the performance of a microstrip antenna as a function of temperature for TM_{01} and TM_{10} is introduced in [16], [17].

III. General Expressions for Electric and Magnetic Fields Intensities

In this section, we will introduce the general expressions of electric and magnetic field intensities for a microstrip antenna printed on a cylindrical body represented in cylindrical coordinates.
Starting from Maxwell's Equation s, we can get the relation between electric field intensity *E* and magnetic flux density *B* as known by Faraday's law [18], as shown in Equation (2):

$$\nabla \times E = -\frac{\partial B}{\partial t} \tag{2}$$

Magnetic field intensity *H* and electric flux density *D* are related by Ampérés law as in Equation (3):

$$\nabla \times H = J + \frac{\partial D}{\partial t} \tag{3}$$

where *J* is the electric current density.
The magnetic flux density *B* and electric flux density *D* as a function of time *t* can be written as in Equation (4):

$$B(t) = \mu\, H\, e^{-j\omega t}$$

and $$D(t) = \varepsilon\, E\, e^{-j\omega t} \tag{4}$$

where μ is the magnetic permeability and ε is the electric permittivity.
By substituting Equation (4) in Equations (2) and (3), we can get:

$$\nabla \times E = -j\omega\mu\mathrm{H}$$

and $$\nabla \times H = j\omega\varepsilon E + J \tag{5}$$

where ω is the angular frequency and has the form of: $\omega = 2\pi f$. In homogeneous medium, the divergence of Equation (2) is:

$$\nabla . H = 0$$

and $$H = \nabla \times A \tag{6}$$

From Equation (5), we can get Equation (7):

$$\nabla \times E + j\omega\mu\mathrm{H} = 0$$

or $$\nabla \times (E + j\omega\mu\mathrm{A}) = 0 \tag{7}$$

Using the fact that, any curl free vector is the gradient of the same scalar, hence:

$$(E + j\omega\mu\mathrm{A}) = -\nabla\varphi \tag{8}$$

where φ is the electric scalar potential.
By letting:

$$\nabla . A = -j\omega\mu\varphi$$

where *A* is the magnetic vector potential.
So, the Helmholtz Equation takes the form of (9):

$$\nabla^2\mathrm{A} + k^2 = \text{-J} \tag{9}$$

k is the wave number and has the form of: $k = \omega\sqrt{\mu\varepsilon}$, and ∇^2 is Laplacian operator. The solutions of Helmholtz Equation are called wave potentials:

$$E = -j\omega\mu\varepsilon A + \frac{1}{j\omega\varepsilon}\nabla(\nabla.\mathrm{A}) \tag{10}$$
$$H = \nabla \times A$$

A) Near Field Equations

By using the Equations number (10) and magnetic vector potential in [19], we can get the near electric and magnetic fields as shown below:

$$E_z = \frac{1}{2\pi j\omega\varepsilon}\sum_{n=-\infty}^{\infty} e^{jn\emptyset}\int_{-\infty}^{\infty}(k^2 - k_z^2) f_n(k_z) H_n^{(2)}\left(\rho\sqrt{k-k_z^2}\right) e^{jk_z z}\, dk_z \tag{12}$$

E_φ and E_ρ are also getting using Equation (7);

$$E_\emptyset = -\frac{1}{2\pi j\omega\varepsilon}\sum_{n=-\infty}^{\infty} e^{jn\emptyset}\int_{-\infty}^{\infty} k_z f_n(k_z) H_n^{(2)}\left(\rho\sqrt{k-k_z^2}\right) e^{jk_z z}\, dk_z \tag{13}$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$E_\rho = \frac{1}{2\pi j\omega\varepsilon}\sum_{n=-\infty}^{\infty} e^{jn\emptyset}\int_{-\infty}^{\infty}\sqrt{k^2-k_z^2}f_n(k_z)H_n^{(2)'}\left(\rho\sqrt{k-k_z^2}\right)e^{jk_zz}\,dk_z \quad (14)$$

To get the magnetic field in all directions, we can use the second part of Equation (10) as shown below, where H_z= 0 for *TM* mode:

$$H_\emptyset = \frac{\partial\psi}{\partial\emptyset} = \frac{j}{2\pi}\sum_{n=-\infty}^{\infty} ne^{jn\emptyset}\int_{-\infty}^{\infty} f_n(k_z)H_n^{(2)}\left(\rho\sqrt{k-k_z^2}\right)e^{jk_zz}\,dk_z \quad (15)$$

$$H_\rho = -\frac{\partial\psi}{\partial\rho} = -\frac{1}{2\pi}\sum_{n=-\infty}^{\infty}\int_{-\infty}^{\infty} f_n(k_z)\sqrt{k^2-k_z^2}H_n^{(2)'}\left(\rho\sqrt{k-k_z^2}\right)e^{jk_zz}\,dk_z \quad (16)$$

B) Far field Equations

In case of far field, we need to represent the electric and magnetic field in terms of *r*, where *r* is the distance from the center to the point that we need to calculate the field on it. By using the cylindrical coordinate Equations, one can notice that a far field ρ tends to infinity when *r*, in Cartesian coordinate, tends to infinity. Also, using simple vector analysis, one can note that, the value of k_z will equal to $-k\times cos\theta$ [19], and from the characteristics of Hankel function, we can rewrite the magnetic vector potential illustrated in Equation (12) to take the form of far field as illustrated in Equation (17).

$$A_z \xrightarrow{r\to\infty} \frac{e^{-jkr}}{\pi r}\sum_{n=-\infty}^{\infty} e^{jn\emptyset}j^{n+1}f_n(-kcos\theta) \quad (17)$$

Hence, the electric and magnetic field can easily be calculated as shown below:

$$E_z = \frac{e^{-jkr}}{j\omega\varepsilon\pi r}k^2\sum_{n=-\infty}^{\infty} e^{jn\emptyset}j^{n+1}f_n(-kcos\theta) \quad (18)$$

$$E_\emptyset = \frac{e^{-jkr}}{j\omega\varepsilon\pi r}\sum_{n=-\infty}^{\infty} jn\,e^{jn\emptyset}j^{n+1}f_n(-kcos\theta) \quad (19)$$

$$E_r = \frac{e^{-jkr}(1+jkr)}{j\omega\varepsilon\pi r^2}\sum_{n=-\infty}^{\infty} e^{jn\emptyset}j^{n+1}f_n(-kcos\theta) \quad (20)$$

The magnetic field intensity also obtained as shown below, where H_z = *0:*

$$H_r = \frac{e^{-jkr}(1+jkr)}{r^2}\sum_{n=-\infty}^{\infty} e^{jn\emptyset}j^{n+1}f_n(-kcos\theta) \quad (21)$$

$$H_\emptyset = \frac{-e^{-jkr}}{\pi r}\sum_{n=-\infty}^{\infty} n\,e^{jn\emptyset}j^{n+2}f_n(-kcos\theta) \quad (22)$$

IV. INPUT IMPEDANCE

The input impedance is defined as "the impedance presented by an antenna at its terminals" or "the ratio of the voltage current at a pair of terminals" or "the ratio of the appropriate components of the electric to magnetic fields at a point". The input impedance is a function of the feeding position as we will see in the next few lines.

To get an expression of input impedance Z_{in} for the cylindrical microstrip antenna, we need to get the electric field at the surface of the patch. In this case, we can get the wave equation as a function of excitation current density *J* as follow:

$$\frac{1}{\rho^2}\frac{\partial^2E_\rho}{\partial\emptyset^2}+\frac{\partial^2E_\rho}{\partial z^2}+k^2E_\rho = j\omega\mu J \quad (23)$$

By solving this Equation, the electric field at the surface can be expressed in terms of various modes of the cavity as [15]:

$$E_\rho(z,\emptyset) = \sum_n\sum_m A_{nm}\psi_{nm}(z,\emptyset) \quad (24)$$

where A_{nm} is the amplitude coefficients corresponding to the field modes. By applying boundary conditions, homogeneous wave Equation and normalized conditions for ψ_{nm}, we can get an expression for ψ_{nm}as shown below:

1. ψ_{nm} vanishes at the both edges for the length *L*:
$$\frac{\partial\psi}{\partial z}\Big|_{z=0} = \frac{\partial\psi}{\partial z}\Big|_{z=L} = 0 \quad (25)$$
2. ψ_{nm} vanishes at the both edges for the width *W*:
$$\frac{\partial\psi}{\partial\emptyset}\Big|_{\emptyset=-\theta_1} = \frac{\partial\psi}{\partial\emptyset}\Big|_{\emptyset=\theta_1} = 0 \quad (26)$$
3. ψ_{nm} should satisfy the homogeneous wave Equation :
$$(\frac{1}{\rho^2}\frac{\partial^2}{\partial\emptyset^2}+\frac{\partial^2}{\partial z^2}+k^2)\psi_{nm} = 0 \quad (27)$$
4. ψ_{nm} should satisfy the normalized condition:
$$\int_{z=0}^{z=L}\int_{\emptyset=-\theta_1}^{\emptyset=\theta_1}\psi_{nm}\,\psi_{nm}^* = 1 \quad (28)$$

Hence, the solution of ψ_{nm}will take the form shown below:

$$\psi_{nm}(z,\emptyset) = \sqrt{\frac{\varepsilon_m\varepsilon_n}{2a\theta_1L}}\cos(\frac{m\pi}{2\theta_1}(\emptyset-\emptyset_1))\cos(\frac{n\pi}{L}z) \quad (29)$$

with

$$\varepsilon_p = \begin{cases}1 & for\ p=0\\ 2 & for\ p\neq 0\end{cases}$$

The coefficient A_{mn} is determined by the excitation current. For this, substitute Equation (29) into Equation (23) and multiply both sides of (23) by ψ_{nm}^*, and integrate over area of the patch. Making use of orthonormal properties of ψ_{nm}, one obtains:

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$A_{nm} = \frac{j\omega\mu}{k^2 - k_{nm}^2} \iint_{dim}^{feed} \psi_{nm}^* J_\rho \, d\emptyset dz \quad (30)$$

Now, let the coaxial feed as a rectangular current source with equivalent cross-sectional area $S_z \times S_\emptyset$ centered at$(Z_0, \emptyset_0)$, so, the current density will satisfy the Equation below:

$$J_\rho = \begin{cases} \frac{I_0}{S_z \times S_\emptyset} & Z_0 - \frac{S_z}{2} \le x \le Z_0 + \frac{S_z}{2} \\ & \emptyset_0 - \frac{S_\emptyset}{2} \le x \le \emptyset_0 + \frac{S_\emptyset}{2} \\ 0 & elsewhere \end{cases} \quad (31)$$

Use of Equation (31) in (30) gives:

$$A_{nm} = \frac{j\omega\mu I}{k^2 - k_{nm}^2} \sqrt{\frac{\varepsilon_m \varepsilon_n}{2a\theta_1 L}} \cos\left(\frac{m\pi}{2\theta_1}\emptyset_0\right) \cos\left(\frac{n\pi}{L}z_0\right) sinc(\frac{n\pi}{2L}z_0) sinc(\frac{m\pi}{2a\theta_1}\emptyset_0) \quad (32)$$

So, to get the input impedance, one can substitute in the following Equation:

$$Z_{in} = \frac{V_{in}}{I_0} \quad (33)$$

where V_{in} is the *RF* voltage at the feed point and defined as:

$$V_{in} = -E_\rho(z_0, \emptyset_0) \times h \quad (34)$$

By using Equations (24), (29), (32), (34) and substitute in (33), we can obtain the input impedance for a rectangular microstrip antenna conformal in a cylindrical body as in the following Equation:

$$Z_{in} = j\omega\mu h \sum_n \sum_m \frac{1}{k^2 - k_{nm}^2} \frac{\varepsilon_m \varepsilon_n}{2a\theta_1 L} \cos^2\left(\frac{m\pi}{2\theta_1}\emptyset_0\right) \cos^2\left(\frac{n\pi}{L}z_0\right) \times sinc(\frac{n\pi}{2L}z_0) sinc(\frac{m\pi}{2a\theta_1}\emptyset_0) \quad (35)$$

V. Voltage Standing Wave Ratio and Return Loss

Voltage Standing Wave Ration *VSWR* is defined as the ration of the maximum to minimum voltage of the antenna. The reflection coefficient ρ define as a ration between incident wave amplitude V_i and reflected voltage wave amplitude V_r, and by using the definition of a voltage reflection coefficient at the input terminals of the antenna Γ, as shown below:

$$\Gamma = \frac{Z_{input} - Z_0}{Z_{input} + Z_0} \quad (36)$$

where, Z_0 is the characteristic impedance of the antenna. If the Equation is solved for the reflection coefficient, it is found that, where the reflection coefficient ρ is the absolute vale of the magnitude of Γ,

$$\rho = |\Gamma| = \frac{VSWR - 1}{VSWR + 1} \quad (37)$$

Consequently,

$$VSWR = \frac{|\Gamma| + 1}{|\Gamma| - 1} \quad (38)$$

The characteristic can be calculated as in [14],

$$Z_0 = \sqrt{\frac{L}{C}} \quad (39)$$

where : L is the inductance of the antenna, and C is the capacitance and can be calculated as follow:

$$C = \frac{2\pi\varepsilon}{\ln\left(\frac{a+h}{a}\right)} \frac{W}{2\pi\varepsilon} \quad (40)$$

$$L = \frac{\mu}{2\pi} \ln\left(\frac{a+h}{a}\right) \frac{W}{2\pi\varepsilon} \quad (41)$$

Hence, we can get the characteristic impedance as shown below:

$$Z_0 = \frac{1}{2\pi} \sqrt{\frac{\mu}{\varepsilon}} \ln\left(\frac{a+h}{a}\right) \quad (42)$$

The return loss s_{11} is related through the following Equation:

$$s_{11} = -20 \log\left[\frac{V_r}{V_i}\right] = -20 \log\left[\frac{VSWR - 1}{VSWR + 1}\right] \quad (43)$$

VI. Results

For the range of GHz, the dominant mode is TM_{01} for $h << W$ which is the case. Also, for the antenna operates at the ranges 2.15 and 1.93 GHz for two different substrates we can use the following dimensions; the original length is 41.5 cm, the width is 50 cm and for different lossy substrate we can get the effect of curvature on the effective dielectric constant and the resonance frequency.

Two different substrate materials RT/duroid-5880 PTFE and K-6098 Teflon/Glass are used for verifying the new model. The dielectric constants for the used materials are 2.2 and 2.5 respectively with a tangent loss 0.0015 and 0.002 respectively.

A) RT/duroid-5880 PTFE Substrate

The mathematical and experimental results for input impedance, real and imaginary parts for a different radius of curvatures are shown in Figures 5 and 6. The peak value of the real part of input impedance is almost 250 Ω at frequency 2.156 GHz which gives a zero value for the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

imaginary part of input impedance as shown in Figure 6 at 20 mm radius of curvature. The value 2.156 GHz represents a resonance frequency for the antenna at 20 mm radius of curvature.

VSWR is given in Figure 7. It is noted that, the value of VSWR is almost 1.4 at frequency 2.156 GHz which is very efficient in manufacturing process. It should be between 1 and 2 for radius of curvature 20 mm. The minimum VSWR we can get, the better performance we can obtain as shown clearly from the definition of VSWR.

Return loss (S11) is illustrated in Figure 8. We obtain a very low return loss, -36 dB, at frequency 2.156 GHz for radius of curvature 20 mm.

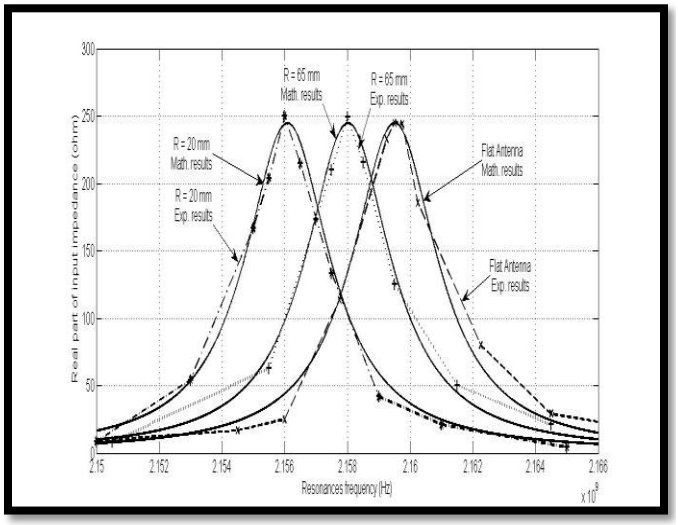


FIGURE 5. Mathimatical and experimental real part of the input impedance as a function of frequency for different radius of curvatures.

Normalized electric field for different radius of curvatures is illustrated in Figure 9. Normalized electric field is plotted for θ from zero to 2π and φ equal to zero. As the radius of curvature is decreasing, the radiated electric field is getting wider, so electric field at 20 mm radius of curvature is wider than 65 mm and 65 mm is wider than flat antenna. Electric field strength is increasing with decreasing the radius of curvature, because a magnitude value of the electric field is depending on the effective dielectric constant and the effective dielectric constant depending on the radius of curvature which decreases with increasing the radius of curvature.

Normalized magnetic field is wider than normalized electric field, and also, it is increasing with deceasing radius of curvature. Obtained results are at for θ from zero to 2π and φ equal to zero and for radius of curvature 20, 65 mm and for a flat microstrip printed antenna are shown in Figure 10. For different radius of curvature, the resonance frequency changes according to the change in curvature, so the given normalized electric and magnetic fields are calculated for different resonance frequency according to radius of curvatures.

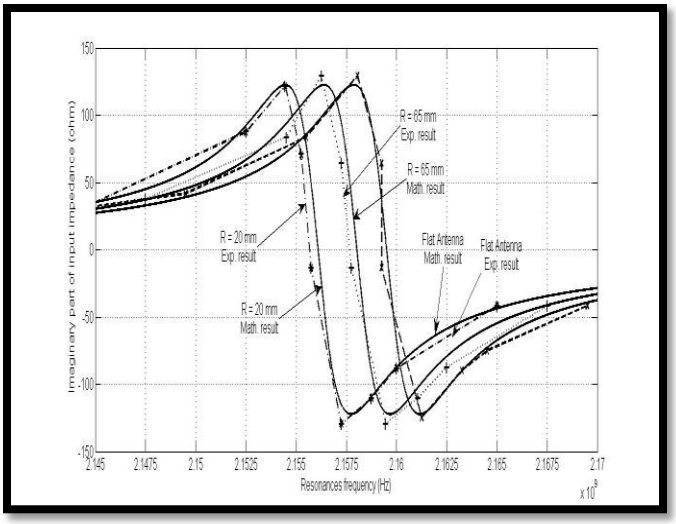


FIGURE 6. Mathimatical and experimental imaginary part of the input impedance as a function of frequency for different radius of curvatures.

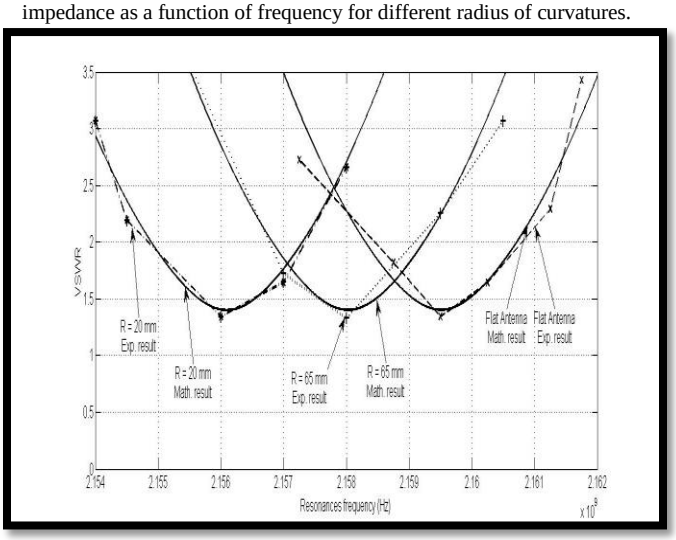


FIGURE 7. Mathimatical and experimental VSWR versus frequency for different radius of curvatures.

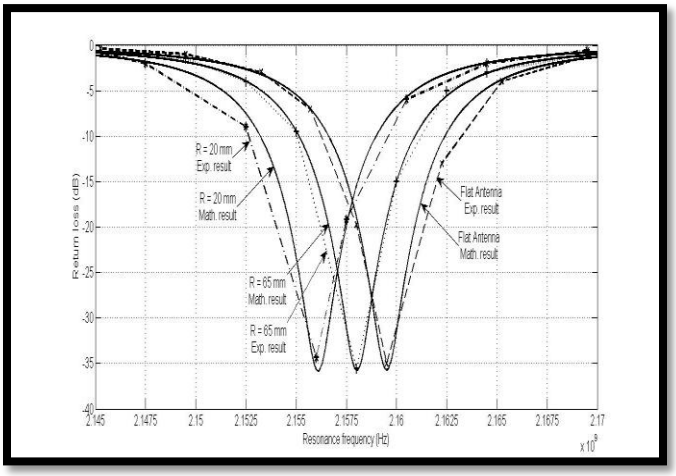


FIGURE 8. Mathimatical and experimental return loss (S11) as a function of frequency for different radius of curvatures.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

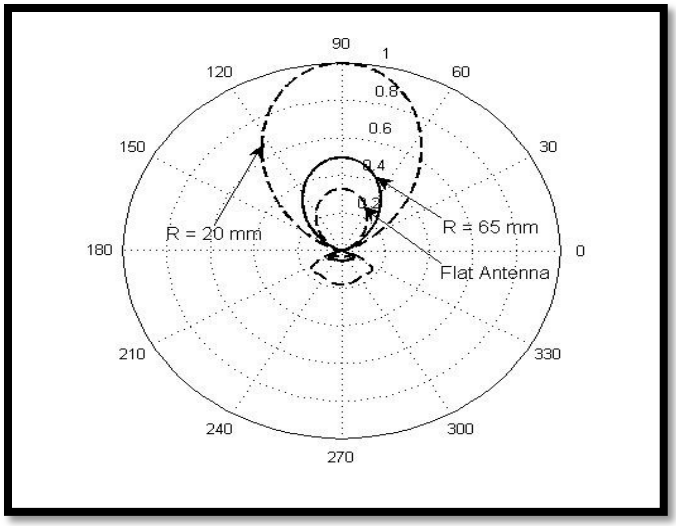


FIGURE 9. Normalized electric field for radius of curvatures 20, 65 mm abd a flat antenna at $\theta=0:2\pi$ and $\varphi=0^0$.

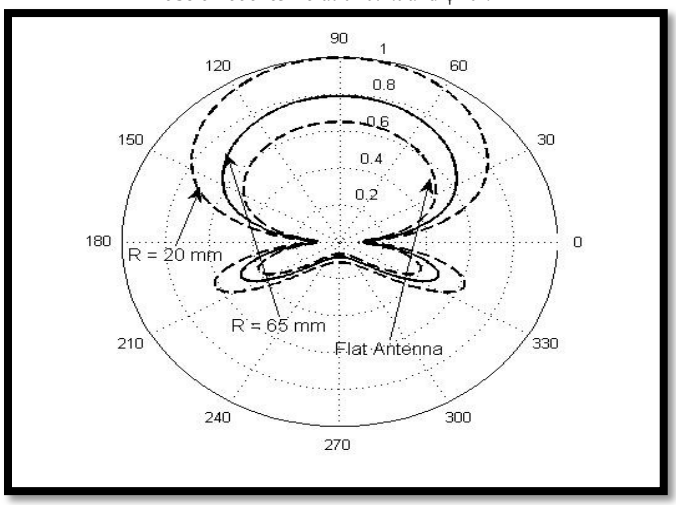


FIGURE 10. Normalized magnetic field for radius of curvatures 20, 65 mm abd a flat antenna at $\theta=0:2\pi$ and $\varphi=0^0$.

B) K-6098 Teflon/Glass Substrate

The real part of input impedance is given in Figure 11 as a function of curvature for 20 and 65 mm radius of curvature compared to a flat microstrip printed antenna. The peak value of a real part of input impedance at 20 mm radius of curvature occurs at frequency 1.935 GHz at 330 Ω maximum value of resistance. The imaginary part of input impedance, Figure 12, is matching with the previous result which gives a zero value at this frequency. The resonance frequency at 20 mm radius of curvature is 1.935 GHz, which gives the lowest value of a VSWR, Figure 13, and lowest value of return loss as in Figure 14. Return loss at this frequency is -50 dB which is a very low value that leads a good performance for a microstrip printed antenna regardless of input impedance at this frequency.

The normalized electric field for K-6098 Teflon/Glass substrate is given in Figure 15 at different radius of curvatures 20, 65 mm and for a flat microstrip printed antenna.

Normalized electric field is calculated at θ equal to values from 0 to 2π and φ equal to zero. At radius of curvature 20 mm, the radiation pattern of normalized electric field is wider than 65 mm and flat antenna, radiation pattern angle is almost 120^0, and gives a high value of electric field strength due to effective dielectric constant.

The normalized magnetic field is given in Figure 16, for the same conditions of normalized electric field. Normalized magnetic field is wider than normalized electric field for 20 mm radius of curvature; it is almost 170^0 for 20 mm radius of curvature. So, for normalized electric and magnetic fields, the angle of transmission is increased as a radius of curvature decreased.

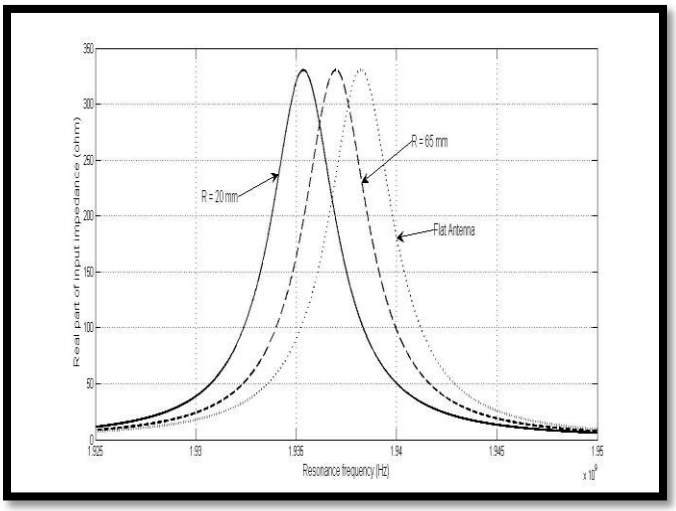


FIGURE 11. Real part of the input impedance as a function of frequency for different radius of curvatures.

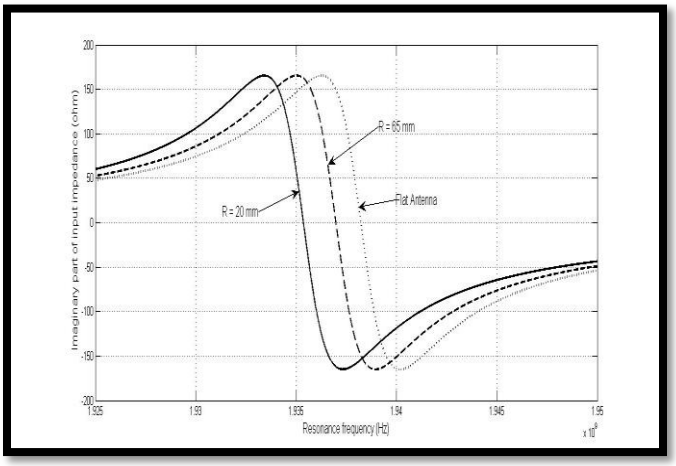


FIGURE 12. Imaginary part of the input impedance as a function of frequency for different radius of curvatures.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

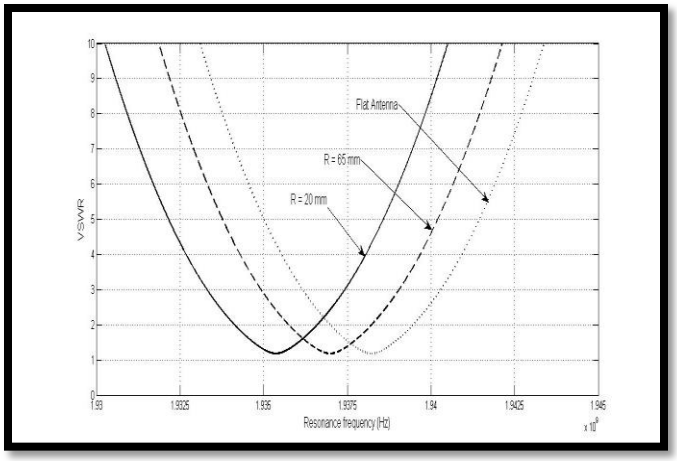


FIGURE 13. VSWR versus frequency for different radius of curvatures.

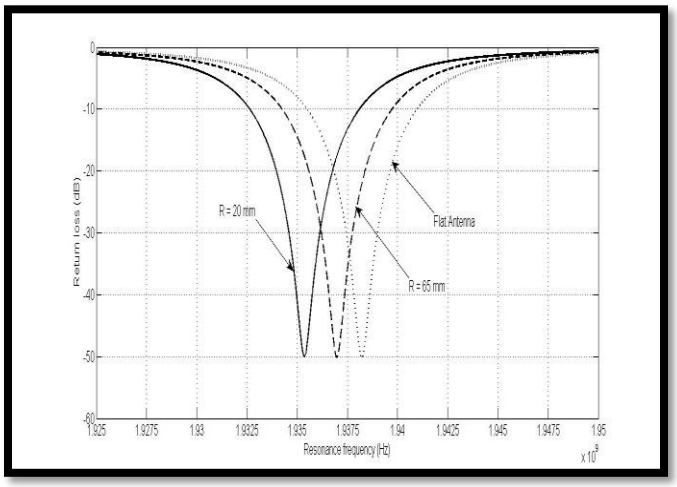


FIGURE 14. Return loss (S11) as a function of frequency for different radius of curvatures.

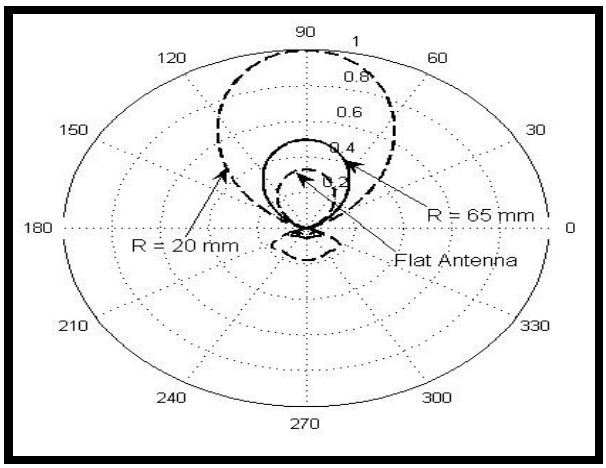


FIGURE 15. Normalized electric field for radius of curvatures 30, 50 and 70 mm at $\theta=0:2\pi$ and $\varphi=0^0$.

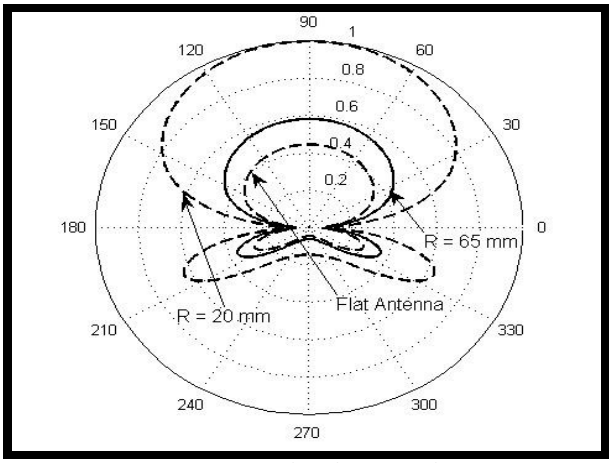


FIGURE 16. Normalized magnetic field for radius of curvatures 20, 65 mm abd a flat antenna at $\theta=0:2\pi$ and $\varphi=0^0$.

CONCLUSION

The effect of curvature on the performance of conformal microstrip antenna on cylindrical bodies for TM_{01} mode is studied in this paper. Curvature affects the fringing field and fringing field affects the antenna parameters. The Equations for real and imaginary parts of input impedance, return loss, VSWR and electric and magnetic fields as a functions of curvature and effective dielectric constant are derived. By using these derived equations, we introduced the results for different dielectric conformal substrates. For the two dielectric substrates, the decreasing in frequency due to increasing in the curvature is the trend for all materials and increasing the radiation pattern for electric and magnetic fields due to increasing in curvature is easily noticed.

We conclude that, increasing the curvature leads to increasing the effective dielectric constant, hence, resonance frequency is increased. So, all parameters are shifted toward increasing the frequency with increasing curvature.

REFERENCES

[1] Heckler, M.V., et al., CAD Package to Design Rectangular Probe-Fed Microstrip Antennas Conformed on Cylindrical Structures. roceedings of the 2003 SBMO/IEEE MTT-S International, Microwave and Optoelectronics Conference, 2003. , 2003. . 2: p. 747-757.

[2] Q. Lu, X. Xu, and M. He, Application of Conformal FDTD Algorithm to Analysis of Conically Conformal Microstrip Antenna. IEEE International Conference on Microwave and Millimeter Wave Technology, ICMMT 2008. , April 2008. 2: p. 527 – 530.

[3] Wong, K.L., Design of Nonplanar Microstrip Antennas and Transmission Lines. 1999: John & Sons, Inc, .

[4] Josefsson, L. and P. Persson, Conformal Array Antenna Theory and Design 1ed. 2006: Wiley-IEEE Press.

[5] Thomas, W., R.C. Hall, and D. I. Wu, Effects of curvature on the fabrication of wraparound antennas IEEE International Symposium on Antennas and Propagation Society,, 1997. 3: p. 1512-1515.

[6] J. Byun, B. Lee, and F.J. Harackiewicz, FDTD Analysis of Mutual Coupling between Microstrip Patch Antennas on Curved Surfaces.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IEEE International Symposium on Antennas and Propagation Society, 1999. 2: p. 886-889.
[7] Balanis, C.A., AntennaTheory. 2005, New York: John Wiley & Sons.
[8] Pozar, D., Microstrip Antennas. IEEE Antennas and Propagation Proceeding, 1992. 80(1).
[9] Krowne, C.M., Cylindrical-Rectangular Microstrip Antenna. IEEE Trans. on Antenna and Propagation, 1983. AP-31: p. 194-199.
[10] Q. Wu, M. Liu, and Z. Feng, A Millimeter Wave Conformal Phased Microstrip Antenna Array on a Cylindrical Surface. IEEE International Symposium on Antennas and Propagation Society, 2008: p. 1-4.
[11] J. Ashkenazy, S. Shtrikman, and D. Treves, Electric Surface Current Model for the Analysis of Microstrip Antennas on Cylindrical Bodies. IEEE Trans. on Antenna and Propagation, 1985. AP-33: p. 295-299.
[12] K. Luk, K. Lee, and J. Dahele, Analysis of the Cylindrical-Rectangular Patch Antenna. IEEE Trans. on Antenna and Propagation, 1989. 37: p. 143-147.
[13] S. Lei, et al., Anti-impact and Over-loading Projectile Conformal Antennas for GPS,. EEE 3rd International Workshop on Signal Design and Its Applications in Communications, 2007: p. 266-269.
[14] Kolev, N.Z., Design of a Microstip Conform GPS Patch Antenna. IEEE 17th International Conference on Applied Electromagnetic and Communications, 2003: p. 201-204.
[15] A. Elrashidi, K. Elleithy, and Hassan Bajwa, "The Fringing Field and Resonance Frequency of Cylindrical Microstrip Printed Antenna as a Function of Curvature," International Journal of Wireless Communications and Networking (IJWCN), Jul.-Dec., 2011.
[16] A. Elrashidi, K. Elleithy, and Hassan Bajwa, "Effect of Temperature on the Performance of a Cylindrical Microstrip Printed Antenna for TM01 Mode Using Different Substrates," International Journal of Computer Networks & Communications (IJCNC), Jul.-Dec., 2011.
[17] A. Elrashidi, K. Elleithy, and Hassan Bajwa, "The Performance of a Cylindrical Microstrip Printed Antenna for TM10 Mode as a Function of Temperature for Different Substrates," International Journal of Next-Generation Networks (IJNGN), Jul.-Dec., 2011.
[18] S. M. Wentworth, Applied Electromagnetics, John Wiley & Sons, Sons, New York, 2005.
[19] R. F. Richards, Time-Harmonic Electromagnetic Fields, New York: McGraw-Hill, 1961.
[20] R. Garg, P. Bhartia, I. Bahl, and A. Ittipiboon, Microstrip Antenna Design Handbook, Aetech House, Boston, 2001

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A Password-Based authentication and Key Agreement Protocol for Wireless LAN Based on Elliptic Curve and Digital Signature

Saed Rezayi
Department of Electrical Engineering
Amir kabir University of Tehran
Tehran, Iran
saed.rezaei@aut.ac.ir

Mona Sotoodeh
Department of Applied Mathematics
Science and Research Azad University
Tehran, Iran
m.sotoodeh@srbiau.ac.ir

Hojjat Esmaili
Department of Computer Engineering
Sharif University of Tehran
hojjat.esmaili@gmail.com

***Abstract*—Password-based authentication protocols are the strongest among all methods which has been proposed through the period that wireless networks have been rapidly growing, and no perfect scheme has been provided for this sensitive technology. The biggest drawback of strong password protocols is IPR (Intellectual Properties Right); hence they have not become standard; SPEKE, SRP, Snapi and AuthA for instance. In this paper we propose a user-friendly, easy to deploy and PKI-free protocol to provide authentication in WLAN. We utilize elliptic curve and digital signature to improve AMP (Authentication via Memorable Password) and apply it for wireless networks as AMP is not patented and strong enough to secure WLAN against almost all possible known attacks.**

Keywords—WLAN, Password-Based Authentication, AMP, Elliptic Curve, Digital Signature.

I. INTRODUCTION

IEEE 802.11 standard was presented in 1997 and as it is becoming more and more prevalent, security in such networks is becoming a challenging issue and is in great demand. Since wireless standard was introduced, a multitude of protocols and RFCs have been proposed to provide authentication mechanism for entities in a WLAN but a few of them have the chance to become a standard regardless of their strengths.

Apart from this, first password-based key exchange protocol, LGSN [1], was introduced in 1989 and many protocols have followed it. In 1992 first verifier-based protocol, A-EKE [2], presented which was one variant of EKE [3] (Encrypted Key Exchange) a symmetric cryptographic authentication and key agreement scheme. Verifier-based means that client possesses a password while server stores its verifier rather than the password. Next attempt to improve password-based protocols was AKE which unlike EKE was based on asymmetric cryptography; SRP [4] and AMP [5] for instance. These protocols need nothing but a password which is a memorable quantity, hence they are simpler and cheaper to deploy compared with PKI-based schemes. Elliptic curve cryptosystem [6, 7] as a powerful mathematical tool has been applied in cryptography in recent years [8, 9, 10]. The security of Elliptic Curve cryptography relies on the discrete logarithm problem (DLP) over the points on an elliptic curve, whereas the hardness of the RSA [11] public-key encryption and signature is based on integer factorization problem. In cryptography, these problems are used over finite fields in number theory [12].

In this paper elliptic curve cryptosystem is combined with AMP to produce a stronger authentication protocol. To complete the authentication process, any mutually agreeable method can be used to verify that their keys match; the security of the resulting protocol is obviously dependent on the choice of this method. For this part we choose the Elliptic Curve analogue of the Digital Signature Algorithm or ECDSA [13] for short.

The remainder of this paper is organized as follows. In section 2 we give a review about authentication and key agreement concept and requirements in wireless LANs. A brief mathematical background of elliptic curve over finite field is presented in section 3. In section 4 our protocol is proposed. Section 5 describes the security and performance analysis of the proposed protocol. Finally, in section 6 the conclusion and future work is provided.

II. WLAN AUTHENTICATION REQUIERMENTS

Authentication is one of five key issues in network security [14] and it verifies users to be who they say they are. Public Key Infrastructure (PKI [15]) is one of the ways to ensure authentication through digital certificates but not only is highly costly and complicated to implement but also it has risks [16]. Thus, a strong password-based method is the primary choice.

The requirements for authentication in wireless networks, regardless of type of method, are categorized as follows. Since EAP [17] is a common framework in

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

wireless security we refer to this standard to gain some points of it.

A. EAP mandatory requirements specified in [17].

- During authentication, a *strong master session key* must be generated.
- The method which is used for wireless networks must provide *mutual authentication*.
- An authentication method must be *resistant to online and offline dictionary attacks.*
- An authentication method must *protect against man-in-the-middle and replay attacks*.

B. Other requirements related to applicability [18].

- Authentication in wireless networks must achieve *flexibility* in order to adapt to the many different profiles. Authentication also needs to be flexible to suit the different security requirements.
- Authentication model in a WLAN should be *scalable*. Scalability in authentication refers to the ability to adapt from small to large (and vice versa) wireless networks and the capacity to support heavy authentication loads.
- It is valuable for an authentication protocol to be *efficient*. Efficiency within an authentication model is a measure of the costs required to manage computation, communication and storage.
- *Ease of implementation* is another crucial issue because authentication is a burden on administrators' shoulders.

In addition there are some desirable characteristics of a key establishment protocol. Key establishment is a process or protocol whereby a shared secret becomes available to two or more parties, for subsequent cryptographic use. Key establishment is subdivided into key transport and key agreement. A key transport protocol or mechanism is a key establishment technique where one party creates or otherwise obtains a secret value, and securely transfers it to the other(s). While a key agreement protocol or mechanism is a key establishment technique in which a shared secret is derived by two (or more) parties as a function of information contributed by, or associated with, each of these, (ideally) such that no party can predetermine the resulting value [19]. In this paper we are dealing with a key agreement protocol.

C. Requirements of a secure key agreement protocol

- *Perfect forward secrecy* which means that revealing the password to an attacker does not help him obtain the session keys of past sessions.
- A protocol is said to be *resistant to a known-key attack* if compromise of past session keys does not allow a passive adversary to compromise future session keys.
- *Zero-knowledge password proof* means that a party A who knows a password, makes a counterpart B convinced that A is who knows the password without revealing any information about the password itself.

III. MATHEMATICAL BACKGROUND

In this section we briefly discuss about elliptic curve over finite fields, digital signature based on elliptic curve and AMP algorithm.

A. Finite Fields

Let p be a prime number. The finite field F_p, called a prime field, is comprised of the set of integers $\{0,1,2,\dots,p-1\}$ with the following arithmetic operations

- Addition: if $a, b \in F_p$, then $a + b = r$, where r is the reminder when $a + b$ is divided by p and $0 \le r \le p-1$. This is known as addition modulo p.
- Multiplication: if $a, b \in F_p$, then $a.b = s$, where s is the reminder when $a.b$ is divided by p and $0 \le s \le p-1$. This is known as multiplication modulo p.
- Inversion: if a is a non-zero element in F_p, the inverse of a modulo p, denoted a^{-1}, is the unique integer $c \in F_p$ for which $a.c = 1$.

B. Elliptic Curve

Let $p > 3$ be an odd prime. An elliptic curve E defined over F_p is an equation of the form

$$y^2 = x^3 + ax + b \qquad (1)$$

Where $a, b \in F_p$ and $4a^3 + 27b^2 \not\equiv 0 \ (mod\ p)$. The set $E(F_p)$ consists of all points (x, y) with $x, y \in F_p$ which satisfies the equation (1), together with a single element denoted $\mathcal{O}$ and called the *point at infinity*.

There is a rule, called the chord-and-tangent rule, for adding two points on an elliptic curve to give a third elliptic curve point. The following algebraic formulas for the sum of two points and the double of a point can be obtained from this rule (for more details refer to [12]).

- For all $P \in E(F_p)$, $P + \mathcal{O} = \mathcal{O} + P = P$
- If $P = (x, y) \in E(F_p)$, then $(x, y) + (x, -y) = \mathcal{O}$. the point $(x, -y)$ is denoted by $-P$ and is called the negative of P.
- Let $P = (x_1, y_1) \in E(F_p)$ and $Q = (x_2, y_2) \in E(F_p)$, where $P \neq \pm Q$. Then $P + Q = (x_3, y_3)$, where

$$x_3 = \left(\frac{y_2 - y_1}{x_2 - x_1}\right)^2 - x_1 - x_2$$

$$y_3 = \left(\frac{y_2 - y_1}{x_2 - x_1}\right)(x_1 - x_3) - y_1$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Let $P = (x_1, y_1) \in E(F_p)$. Then $2P = (x_3, y_3)$ where

$$x_3 = \left(\frac{3{x_1}^2 + a}{2y_1}\right)^2 - 2x_1$$

$$y_3 = \left(\frac{3{x_1}^2 + a}{2y_1}\right)(x_1 - x_3) - y_1$$

Observe that the addition of two elliptic curve points in $E(F_p)$ requires a few arithmetic operations (addition, subtraction, multiplication, and inversion) in the underlying field.

In many ways elliptic curves are natural analogs of multiplicative groups of fields in Discrete Logarithm Problem (DLP). But they have the advantage that one has more flexibility in choosing an elliptic curve than a finite field. Besides, since the ECDLP appears to be significantly harder than the DLP, the strength-per-key-bit is substantially greater in elliptic curve systems than in conventional discrete logarithm systems. Thus, smaller parameters can be used in ECC than with DL systems but with equivalent levels of security. The advantages that can be gained from smaller parameters include speed (faster computations) and smaller keys. These advantages are especially important in environments where processing power, storage space, bandwidth, or power consumption is constrained like WLANs.

C. AMP

AMP is considered as strong and secure password based authentication and key agreement protocol and is based on asymmetric cryptosystem, in addition, it provides password file protection against server file compromise. Security of AMP is based on two familiar hard problems which are believed infeasible to solve in polynomial time. One is Discrete Logarithm Problem; given a prime p, a generator g of a multiplicative group $\mathcal{Z}_p$, and an element $g^x \in \mathcal{Z}_p$, find the integer $x \in [0,\ p-2]$. The other is Diffie-Hellman Problem [20]; given a prime p, a generator g of a multiplicative group $\mathcal{Z}_p$, and elements $g^x,\ g^y \in \mathcal{Z}_p$, find $g^{xy} \in \mathcal{Z}_p$.

The following notation is used to describe this algorithm according to [13].

id	Entity identification
π	A's password
τ	Password salt
x	A's private key randomly selected from $\mathcal{Z}_p$
y	B's private key randomly selected from $\mathcal{Z}_p$
g	A generator of $\mathcal{Z}_p$ selected by A
$h_i()$	Secure hash functions

AMP^n four pass protocol:

$A\ (id,\pi)$		$B\ (id, g^\pi)$
$x \in \mathcal{Z}_p$		
$G_1 = g^x$	$\xrightarrow{id, g^x}$	
		fetch (id,π)
		$y \in \mathcal{Z}_p$
$w = (x+\pi)^{-1}x$	$\xleftarrow{g^{(x+\pi)y}}$	$G_2 = (G_1 g^\pi)^y$
$\alpha = (G_2)^w$		$\beta = (G_1)^y$
$\mathcal{K}_1 = h_1(\alpha)$		$\mathcal{K}_2 = h_1(\beta)$
$\mathcal{H}_{11} = h_2(G_1, \mathcal{K}_1)$	$\xrightarrow{\mathcal{H}_{11}}$	$\mathcal{H}_{12} = h_2(G_1, \mathcal{K}_2)$
		verify $\mathcal{H}_{11} = \mathcal{H}_{12}$
$\mathcal{H}_{21} = h_3(G_2, \mathcal{K}_1)$	$\xrightarrow{\mathcal{H}_{22}}$	$\mathcal{H}_{22} = h_3(G_2, \mathcal{K}_2)$
verify $\mathcal{H}_{21} = \mathcal{H}_{22}$		

If instead of password, its verifier was stored in server, it would be resistant against server impersonation attack; but we just presented AMP naked in this section. For other variants of AMP refer to [6]. Note that A and B agree on g^{xy}.

D. ECDSA

ECDSA is the elliptic curve variant of DSA which is digital signature mechanism which provides a high level of assurance. There are three main phases in this algorithm; key pair generation, signature generation and signature validation.

Key generation: each entity does the following for domain parameter and associated key pair generation.

1. Select coefficients a and b from F_p verifiably at random. Let E be the curve $y^2 = x^3 + ax + b$.
2. Compute $N = \#E(F_q)$ and verify that N is divisible by a large prime n ($n > 2^{160}$ and $n > 4\sqrt{p}$).
3. Select a random or pseudorandom integer d in the interval $[1, n-1]$.
4. Compute $Q = dG$.
5. The public key is Q; the private key is d.

To assure that a set $D = (p, a, b, G, n)$ of EC domain parameters is valid see [13].

Signature generation: to sign a message m, an entity A with domain parameters D and associated key pair (d, Q) does the following.

1. Select a random or pseudorandom integer k in the interval $[1, n-1]$.
2. Compute $kG = (x_1, y_1)$ and put $r = x_1 \ mod\ n$ if $r = 0$ go to step 1.
3. Compute $e = H(m)$ where H is a strong one way hash function is.
4. Compute $s = k^{-1}(e + dr)\ mod\ n$. If $s = 0$ go to step 1.
5. A's signature for the message m is (r, s).

Signature validation: to verify A's signature on m, B obtains an authentic copy of A's domain parameters D and associated public key Q.

1. Compute $e = H(m)$.
2. Compute $w = s^{-1}\ mod\ n$.
3. Compute $u_1 = ew\ mod\ n$ and $u_2 = rw\ mod\ n$
4. Compute $X = u_1 G + u_2 Q$
5. If $X = \mathcal{O}$, then reject the signature. Otherwise, compute x-coordinnate of X; x_2.
6. Accept the signature if and only if $r = x_2$.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

VI. PROPOSED PROTOCOL

In this section we present our method to improve AMP scheme. As previously mentioned we combine AMP with Elliptic Curve, since smaller parameters can be used in ECC compared with RSA. Besides, the level of latency is quite high in RSA as compared to ECC for the same level of security and for the same type of operations; sign, verification, encryption and decryption. In [21] a key establishment protocol was tested by both ECC and RSA and the latency in millisecond measured as a performance parameters. It is seen from Fig. 1 that RSA has at least four times greater latency than ECC.

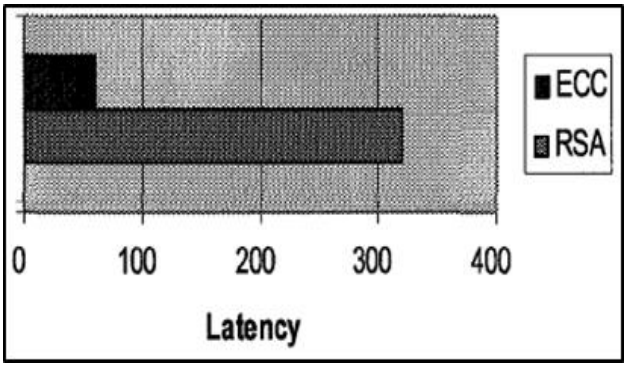


Figure 1: Latency: ECC vs. RSA

Furthermore, for the two last steps, we utilize ECDSA which is a high secure signing method than hash functions. Before running the protocol, entity A chooses an elliptic curve $(\text{i.e. } E(F_p) \text{ over } F_p)$, and then he randomly selects a large prime G from F_p. Moreover (d, Q) is his key pair. We assume that A and B securely shared password π. See section 2 for parameter selection. The rest of the protocol is illustrated as follows.

$A\ (id,\pi)$		$B\ (id, g^{\pi})$
$x \in F_p$		
$X = xG = (x_1, y_1)$	$\xrightarrow{Q,id,X,G}$	
$r = x_1$		*fetch* (id,π)
		$y \in F_p$
$w = (x+\pi)^{-1}$	$\overset{Y}{\leftarrow}$	$Y = y(X + \pi G)$
$S = xwY$		$S = yX$
$e = h(S)$		
$s = x^{-1}(e + dr)$	$\overset{s,r}{\rightarrow}$	$h(S) = e$
		$z = s^{-1}$
		$u_1 = ez\ , u_2 = rz$
		$u_1G + u_2Q = (x_2, y_2)$
		verify $r = x_2$

A randomly selects x from F_p and computes $X = xG = (x_1, y_1)$ and puts $r = x_1$. He sends $X, G,\ Q$ (his public key) and his id to B

1. Upon receiving A's id, B fetches A's password according to received id and randomly selects y, computes $Y = y(X + \pi Q)$, and sends it to A.
2. A computes $w = (x+\pi)^{-1}$ and obtains the session key as follows.

$$\begin{aligned} S = xwY &= x(x+\pi)^{-1}y(X + \pi Q) \\ &= x(x+\pi)^{-1}y(xG + \pi G) \\ &= x(x+\pi)^{-1}y(x+\pi)G = xyG \end{aligned}$$

He signs it as described in section 3.4, and sends (r, s) as digital signature.

3. B also computes the session key as follows.
$S = yX = xyG$
And verifies the validity of digital signature as below,
$z = s^{-1}$
$$= x(e + dr)^{-1}$$
$$\Rightarrow u_1 = ex(e + dr)^{-1}\ ,\ u_2 = rx(e + dr)^{-1}$$
To $r = x_2$ get satisfied following equation must be true:
$$u_1G + u_2Q = xG$$
$$u_1G + u_2Q = ex(e + dr)^{-1}G + rx(e + dr)^{-1}Q$$
$$Q = dG \xrightarrow{yields} (e + dr)^{-1}(e + rd)xG = xG$$

V. SECURITY AND PERFORMANCE ANALYSIS

A. *Security Analysis*

We claim that our proposed protocol is secure enough to be used in sensitive Wireless LANs and protect these networks against well-known attacks. Because the security of the authentication model depends on the security of the individual protocols in the model; AMP and ECDSA, besides more flexible and stronger cryptosystem is applied to make it applicable in WLANs. In addition to generating strong session key and providing mutual authentication, following properties are presented to prove our protocol strength.

Perfect Forward Secrecy: our protocol provides perfect forward secrecy (as AMP and other strong password based protocols do) via Diffie-Hellman problem and DLP and due to the complicacy of these problems. Because even if an adversary eavesdrops π, he cannot obtain old session keys because the session key is formed by random numbers, x and y, generated by both entities which are not available and obtainable.

Man in the Middle Attack: this attack is infeasible because an attacker does not know the password π. Assume he is in the middle of traffic exchange and A, B have no idea about this. He gets A's information but does not send them to B, instead, he stores them and selects a large prime from F_p, let k, then he computes $K = kG$ and sends it to B. B computes $Y = y(K + \pi G)$ and sends it to A. on the way, attacker grabs Y and sends it to A, but A and B shared session key, S, does not match due to wrong digital signature which A produced.

Dictionary Attack: offline dictionary attack is not feasible because an adversary, who guesses the password π, has to solve DLP problem to find y in equation $Y = y(X + \pi G)$ and obtains S. Online dictionary attack is also

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

not applicable because the entity A is never asked for password.

Replay Attack: is negligible because *X* should include an ephemeral parameter of A while *Y* should include ephemeral parameters of both parties of the session. Finding those parameters corresponds to solving the discrete logarithm problem.

Zero Knowledge Password Proof: this property is provided since no information about password is exchanged between two parties.

Known-Key Attack: our protocol resists this attack since session keys are generated by random values which are irrelevant in different runs of protocol.

B. Performance Analysis

Flexibility: our protocol is based on AMP, and AMP has several variants for various functional considerations. So it can implemented in every scenarios; wired or wireless. For example, as we mentioned, one variant of AMP is secure against password-file compromise attack whereas another is useful for situations where are very restricted and A, B are allowed to send only one message.

Scalability: since AMP has light constraints and is easy to generalize and because of its low management costs and low administrative overhead unlike PKI, our proposed protocol is highly scalable.

Efficiency: AMP is the most efficient protocol among the existing verifier-based protocols regarding several factors such as the number of protocol steps, large message blocks and exponentiations [6]. Hence a generalization of AMP on elliptic curve is very useful for further efficiency in space and speed.

Ease of Implementation: due to all reasons provided in this sub-section and since our protocol does not need any particular Infrastructure, it can be implemented easily.

VI. CONCLUSION AND FUTURE WORK

In this work we proposed a password-based authentication and key agreement protocol based on elliptic curve for WLAN. In fact we modified AMP and applied ECDSA digital signature standard to amplify the security of AMP since elliptic curve cryptosystem is stronger and more flexible. Further, we showed that our protocol has all parameters related to security and applicability. Besides, it satisfies all mandatory requirements of EAP.

For future work a key management scheme can be designed and placed in layering model to manage and refresh keys for preventing cryptanalysis attacks. Besides, this protocol can be implemented in OPNET simulator to gain advantages from more statistical parameters and it can be compared with other authentication protocols using OPNET.

REFRENCES

[1] M. Lomas, L. Gong, J. Saltzer, and R. Needham, "Reducing risks from poorly chosen keys," *ACM Symposium on Operating System Principles*, 1989, pp.14-18.

[2] S. Bellovin and M. Merritt, "Augmented encrypted key exchange: a password-based protocol secure against dictionary attacks and password-file compromise," *Proceedings of the 1st ACM Conference on Computer and Communications Security*, 1993, pp. 244-250.

[3] S. Bellovin and M. Merritt, "Encrypted key exchange: password-based protocols secure against dictionary attacks," *Proc. IEEE Comp. Society Symp. on Research in Security and Privacy*, 1992, pp. 72-84.

[4] T. Wu, "Secure remote password protocol," *Internet Society Symposium on Network and Distributed System Security*, 1998.

[5] T. Kwon, "Authentication and key agreement via memorable passwords," *In Proceedings of the ISOC Network and Distributed System Security (NDSS)*, 2001.

[6] V. Miller, "Uses of elliptic curves in cryptography", *Advances in Cryptology, Lecture Notes in Computer Science, Springer-Verlag*, 1986, pp. 417-426.

[7] N. Koblitz, "Elliptic curve cryptosystems", *Mathematics of Computation*, 1987, pp. 203-209.

[8] C. Tang, and D. O. Wu, "An Efficient Mobile Authentication Scheme for wireless networks," *IEEE Transactions on Wireless Communications, Vol. 7, No. 4*, 2008, pp. 1408-1416.

[9] H. Zhu, and T. Liu, "A Robust and Efficient Password-authenticated key agreement scheme without verification table Based on elliptic curve cryptosystem," *International Conference on Computational Aspects of Social Networks*, 2010, pp. 74-77.

[10] K. R. Pillai, and M. P. Sebastian, "Elliptic Curve based Authenticated Session Key Establishment Protocol for High Security Applications in Constrained Network Environment," *International Journal of Network Security & Its Applications (IJNSA), Vol.2, No.3*, 2010, pp. 144-156.

[11] R Rivest, A. Shamir, and L. Adleman, "A Method for Obtaining Digital Signatures and Public Key Crypto-systems," *Communications of the ACM, Vol. 21, No. 2*, 1978.

[12] N. Koblitz, *A Course in Number Theory and Cryptography*, 2nd edition, Springer-Verlag, 1994.

[13] D. Johnson, A. Menezes, and S. Vanstone, "The Elliptic Curve Digital Signature Algorithm (ECDSA)," *International Journal of Information Security, Vol. 1, No. 1*, 2001 pp. 36-63.

[14] W. Peterson, and C. Scott, *Tactical Perimeter Defense*, Security Certified Program, LLC, 2007.

[15] R. Housley, and T. Polk, *Planning for PKI*, John Wiley & Sons, New York, 2001.

[16] C. Ellison, and B. Schneier, "Ten Risks of PKI: What You Are not Being Told about Public Key Infrastructure," Computer Security Journal, Vol. 17, No. 1, 2000.

[17] B. Aboba, L. Blunk, J. Vollbrecht, J. Carlsonand, and H. Levkowetz, RFC 3748 "Extensible Authentication Protocol (EAP)," June 2004 [Online]. Available: http://tools.ietf.org/html/rfc3748.

[18] H. H. Ngo, "Dynamic Group-Based Authentication in Wireless Networks," Ph.D. dissertation, Dept. Information Technology, Univ. Monash, 2010.

[19] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*, 1st edition, CRC Press, 1996.

[20] W. Diffie, and M. E. Hellman, "New Directions in Cryptography," *IEEE Transaction on Information Theory, Vol.22, No. 6*, 1996, pp. 644-654.

[21] V. Sethi, and B. Thuraisingham, "A Comparative Study of A Key Agreement Protocol Based on ECC and RSA," Department of Computer Science, The University of Texas at Dallas, Tech. Rep. UTDCS-60-06, Nov. 2006.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Computer Based Information System Functions for Decision Makers in Organizations.

Mohammed Suliman Al-Shakkah*
School of Computing, College of Arts and Sciences
University Utara Malaysia, UUM
06010 UUM - Sintok, Kedah, Malaysia
alshakkah_11@yahoo.com
alshakkah@gmail.com

Wan Rozaini Sheik Osman
School of Computing, College of Arts and Sciences
University Utara Malaysia, UUM
06010 UUM - Sintok, Kedah, Malaysia
rozai174@uum.edu.my

***Abstract*— Computer Based Information System (CBIS) was discussed by many scholars. In this paper a review was conducted for the CBIS types from different point views` scholars. CBIS is important for decision makers (managers) to make decisions at their different levels. Eighteen managers from five organizations were interviewed with structural interviews. The findings showed that only six managers with 33% only are using CBIS in decision making process (DMP). Thus, this indicates the need for future research in Jordan to find out, why CBIS is still not fully adopted by decision makers.**

Keywords- Computer Based Information System, CBIS, Components, Types, Decision making, Manager, Interview.

I. INTRODUCTION

Due to changing environment for organizations, competition, convergence, networked, and costs. Levels of decision makers decreased in flatted organizations. In this paper the researchers want to know how the Computer Based Information System (CBIS) plays a role. CBIS which is an information system that uses computers (automated-IS), consists of: hardware, software, databases, people, telecommunications and procedures, configured to collect, manipulate, store, and process data into information become so important and highly needed [1, 2]. Most types of work require a high number of people, time and effort to accomplish. All jobs that were done manually a century ago have now become easier to do, as a lot of time and cost are now saved with the development of technology. Similarly, seeking data and information especially from manual reports and studies is tedious to scan through to find the necessary information. Thus, to solve the problem and to find a suitable solution, in particular for an urgent issue could take a very long time. Later, organizing and indexing were introduced to help to retrieve these reports easily. With the advancement in technology, huge information could be organized very well and easily referred to whenever required. The information system can be categorized into two groups: (1) manual systems: the old style that deals with papers and reports, (2) automated systems: where computerizing system is used. There are many types of CBIS, where the transaction processing system (TPS) is the system used at the operations level of organizations for routine process. TPS was introduced in 1950 to support the sudden and unexpected needs, hence, CBIS was required in many organizational levels such as management information system (MIS), decision support system (DSS), group decision support system (GDSS), expert system (ES), office information system (OIS), executive information system (EIS), and intelligence organizational information system (IOIS) [3, 4]. Another way of classification described by Mentzas on the CBIS activities which is: (1) Information reporting where the best example here is MIS, (2) communication and negotiation activities (GDSS), and (3) decision activities (DSS, ES), which support selection from the available alternatives, which is the main focus of this research on decision making [3].

CBIS which is information processing systems have components as follows: hardware, software, data, people, and

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

procedures. These components are organized for specific purposes [5].

This paper will answer the following two questions:

Q1: What are the roles (functions) of CBIS in decision making in organizations?

Q2: Are the CBIS used in the Jordanian organizations by their decision makers?

II. PREVIOUS WORK

Scholars looked for the components and types of CBIS from different perspectives as follows:

In 1985, according to [6], the users of CBIS must have common knowledge of such systems. Due to the fact that computers have become more available and much easier to use, this flexibility helps in getting information that is needed, the components of CBIS viewed are: hardware, software, data, models, procedures, and users. In addition, the CBIS consists of four components: hardware; software; people, and data storage. The purpose of CBIS as an information system with computers was used to store and process data in 1988 [7]. Also, in 1987 and referring to [8], the problem of end-users contributed to the lack of success in the integration of the CBIS system of the organizations. Hence, they presented a quick and powerful solution by means of training the end users to use the IT (CBIS) system. After analyzing several different types of organizational conflicts, in 1990 scholars as [9] suggested that the group decision support system (GDSS) is an essential tool to resolve conflicts. They also perceived that CBIS has evolved from focusing data such as TPS, information such as MIS and decision such as GDSS and DSS. Hence, CBIS and its components are necessary in supporting decision.

In 1994, the components of information processing systems were noted as follows: hardware, software, data, people, and procedures. These components are organized for specific purposes, Furthermore, the researcher mentioned five types of CBIS, from the oldest to the newest, or from more structured to less structure such as; transaction processing systems (TPS), management information systems (MIS), decision support systems (DSS), expert systems (ES) as major type of artificial intelligence (AI) and executive information systems (EIS). Transforming process for data can be classified into three steps such as converting data into information (refining), converting information into decision (interpreting), and installing decisions and changes in the organization (implementing) with some tools as word processing report [5].

In 1995, CBIS was more valuable for manager`s mental model for guiding planning, controlling, and operating decisions, than forming or revising the manager`s mental model of the corporation. The researchers also added that the tools in several studies have shown the most used computer softwares which were spreadsheets, word-processing and data base management. The amount of use was from 1.8 Hr per week to 14Hr or more per week. The lowest use was in Saudi Arabia, while the highest use rate was in Taiwan [10].

However, in 1994 [3] mentioned that from specific types of CBIS (e.g. DSS, GDSS, ES) are powerful tools in certain aspects of the decision making process in the modern organizations, but they have limitations. For example, none of them provide an integrated support. The researcher also made comparison between the ten types of CBIS (MIS, EIS, ESS, DSS, GDSS, EMS ODSS, ES, OIS, and IOIS) to establish and promote for using the IOIS system in organizations. For the roles of these types of CBIS see Table 1.

TABLE 1. TYPES OF COMPUTER-BASED INFORMATION SYSTEM.

Types of CBIS System	Roles of CBIS Types
Management Information System (MIS)	Analysis of information, generation of requested reports, solving of structured problems.
Executive Information System (EIS)	Evaluation of information in timely information analysis for top-level managerial levels in an intelligent manner.
Executive Support Systems (ESS)	Extension of EIS capabilities to include support for electronic communications and organizing facilities.
Decision Support System (DSS)	Use of data, models and decision aids in the analysis of semi-structured problems for individuals.
Group Decision Support System (GDSS)	Extension of DSS with negotiation and communication facilities for group.
Electronic Meeting Systems (EMS)	Provision of information systems infrastructure to support group work and the activities of participants in meetings
Organizational Decision Support Systems (ODSS)	Support of organizational tasks or decision-making activities that affect several organizational units
Expert systems (ES)	Capturing and organizing corporate knowledge about an application domain and translating it into expert advice.
Office Information System (OIS)	Support of the office worker in the effective and timely management of office objects. The goal-oriented and ill-defined office processes and the control of information flow in the office.
Intelligence Organizational Information System (IOIS)	Assistance (and independent action) in all phases of decision making and support in multi participant organizations.

Source: Mentzas (1994).

Mentzas promoted the using of IOIS, and considered it as a perfect solution for supporting decisions in organizations, which was the only type of CBIS that give a high support in three dimensions to (individuals, groups and organizations) as an integration support which is not available in the other nine types mentioned earlier[3].

In 1997, the types of CBIS were in five subsystems comprising data processing (DP), office automation (OA), expert system (ES), decision support system (DSS), and management information system (MIS). Whereas, the researcher promoted for the MIS type to solve the problem in decisions of organizations [11]. In the beginning of this Century (in 2003), the CBIS was considered a vital tool for managers in making decisions. They also, encouraged CBIS courses to be given to the undergraduate students in business administration (BA) in the U.S system through the second year to help them in future. In addition, some of the benefits of CBIS include learning the system design and analysis and improving the problem solving skills [12].

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

In the same year 2003 and according to [4], the CBIS is one unit in which a computer plays the basic role. She presented five components for the CBIS systems namely: Hardware which refers to machines part with input, storages and output parts, software which is a computer programs that helps in processing data to useful information, data in which facts are used by programs to produce useful information, procedures which are the rules for the operations of a computer system, and people or users for the CBIS which are also called end users.

In 2004, scholars as: Vlahos, Ferrat, and Knoepfle found that the CBIS were accepted i.e. (adopted and used) by German managers. Besides, results from their survey have shown that those managers were heavily CBIS users with more than 10 Hr per week. The researchers encouraged using the CBIS system as: it helps in planning, assisting in decision making budgeting, forecasting, and solving problems. As researchers wanted to know how German managers use the CBIS systems, they built a survey questionnaire to collect data. Likert scale with 7-point scale was used; whereas, Cornbach Alpha was 0.77. This study provides a new updated knowledge on CBIS use by German managers, together with looking into the perceived value and satisfaction obtained from CBIS, in helping managers and normal users and supporting them to carry out better decision making [13]

In 2005, according to [14], many decision makers have lack of knowledge in using the automated CBIS. They gave an example where a corporate chief executive has to learn how to use an automated CBIS while his senior managers have limited computer knowledge and so they prefer only extremely easy to use systems. This scenario shows that decision makers want to learn how to use the CBIS to process better decision but they do not know how. In the same year, some scholars as [15] used the term CBIS and IS interchangeably. He also argued for the success of CBIS so as to gain benefits by using information systems (IS) and information technology (IT) in organizations. There is a need to deal with the important needed information with the CBIS to support decision makers.

from the two different years, in 2007 and 2011, Turban, Aronson, Liang, and Sharda decided that the CBIS are required to support decisions in organizations for many reasons such as: works in organizations to rapidly change because of the economy needs to follow the case with the automated systems, to support the decision making process and to have accurate information as required, the management mandates the computerized decision support, high quality of decision is required, the company prefers improved communication and customer and employee satisfaction; timely information is necessary, the organization seeks cost reduction, the organization wants improved productivity, and the information system department of the organization is usually too busy to address all the management's inquiries [16, 17].

In 2007, scholars as [18], noticed that many types of CBIS developed to support decision making are: decision support systems (DSS), group decision support systems (GDSS) and executive information systems (EIS). In their study, they used IS interchangeably with CBIS, and discussed the difference between USA and other Asian countries holding that success depends on how well IT (CBIS) application is adapted to the decision style of their users.

In 2008, a recommendation was by [19], to look for the recommendation systems which are another face for CBIS to support decisions. In his study, he focused on decision DSS, and how they evolved from aiding decision makers to perform analysis to provide automated intelligent support.

In 2009, a promotion to adopt and use after well-understanding of the ICT- in the meaning of CBIS- sector support to give support for the decision making processing by discussing the ICT environment in industrial house construction for six Swedish companies. The interest here was in processing data in a systematic way as organizing the resources for collecting, storage, process, and display information. In these six companies, different ICT support decision tools as (ERP, CAD, Excel, and VB-Scripts seawares) were used. Organizations which did not use ERP system had problems in information management. Again, using ICT models with automated systems (tools) will be a good way to systemize information to reduce cost and save time for the decision makers [20]. In the same year also (2009), scholars as [21] argued that the combinations of two types of CBIS as (DSS with ES) will be a guidance in the process of grading wool for the decision makers in this field. They also added that the DSS has the following advantages. DSS supports decision making activities for the area businesses and organizations, designed to help decision-makers to get useful information after processing raw data. DSS which is an interactive CBIS system was developed to support solving unstructured problems to improve decision-making. Moreover, DSS uses intelligent agents to collect data related to online as auctions which improve decision-making and lastly DSS utilizes statistical analyses that provide the specific and relevant information. In addition, combining DSS with ES will complement the two systems and help decision makers in the decision making process. This will be carried out through a systematic way and will not replace humans as decision makers by the machine or any complex systems.

In 2009, other scholars as [22] argued that it is good to integrate the decision support systems (DSS) which is one type of the CBIS as IDSS as a development system. They discussed more than 100 papers and software systems, and recommended that IDSS will be a better support for decision makers in the decision making process. By looking at literature review, integration of DSS as a tool for users` decision makers was On-Line Analytical Processing (OLAP) as a powerful tool that helps decision makers in processing decisions. Also, in 2009, Fogarty and Armostrong surveyed 171 organizations in Australia for the CBIS or the Automated- IS success which is important for organizations in small business sector and a model for the following factors: organization characteristics, the Chief Executive Officer (CEO) characteristics, decision (Decision Criteria), and user satisfaction. They used the term "small business" to mean a "small and medium enterprise" (SME). This calls for more attention and interest in computer based information systems (CBIS) in organizations to help in the decision making process [23].

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Management support systems (MSS) which is another face for CBIS support different managerial roles i.e. the development of MSS that supports managerial cognition, decision, and action. While CBIS types include: decision support systems DSS), group support systems (GSS), executive information systems (EIS), knowledge management systems (KMS), and business intelligence (BI) systems developed to support the decision making process for managers. On the other hand, MSS have other features such as modeling capabilities, electronic communications, and organizing tools. The researchers here refer to the MSS system as ICT-enabled IS in order to support managers to process decisions which was in 2009 by [24].

In 2010, a comparison by [25] for the traditional-IS with automated-IS (CBIS) system, where they referred to the CBIS system as information system auditing that gives support to the decision makers in their businesses. Computer-based information system is expected to help businesses achieve their goals and objectives, and to lend support for making good decisions by decision makers. They refer to the components of CBIS such as: hardware, software, database, networks, procedures, and people. In the same view, also in the same year (2010), [26] argued that automated system of Customer Relationship Management (CRM) will help not only in the decision making process, but also in reducing costs, and time. In addition, CRM known as software which helps in integration of resources, also helps in sharing knowledge between customers, supports daily decisions, and improves the users` performance.

Other scholars in the same year (2010) as [2], declared that there is a need for CBIS:

"High quality, up-to-date, and well maintained computer-based information systems (CBIS) since they are the heart of today`s most successful corporations" (P. 3).

In addition, they gather the components for CBIS system as a single set of hardware, software, database, telecommunications, people and procedures. They also identified the major role software tool of CBIS which consists of input, processing output, and feedback. The aim is to collect and process data to provide users as decision makers with needed information to help them in the decision making process. One of the examples they gave was SAP software.

In 2010 also, the CBIS can be used to help in industrial process-plants which are important for the economy. A proposed model for determining the financial losses resulting from cyber attacks on CBIS systems was used. The CBIS system here was Supervisory Control and Data Acquisition (SCADA) system. Managers using the SCADA system were helped with estimation about their financial damages. Here, the researchers focus on the risk, cost, resources, and benefits as factors from the decision making to interest with using the CBIS (SCADA) by decision makers [27].

To sum up, the previous components of CBIS, Please, see the following in Table. 2.

TABLE 2. CBIS COMPONENTS.

CBIS components	Researchers
Hardware	[1, 2, 3, 4, 5, 6 & 7]
Software	[1, 2, 4, 5, 6 & 7]
Data storages	[1, 2, 4, 5, 6 & 7]
Models	[3, 6]
Procedure	[1, 2, 4, 5 & 6]
Users	[1, 2, 4, 5, 6 & 7]
Knowledge	[3]
Cooperation	[3]
Support Man-Machine Interaction	[3]
Telecommunications	[1, 2]

In light of the previous discussion, researchers considered the components of CBIS from different points of view with emphasis on, the integration of all to be presented as hardware, software, people, data storage, model and procedures. Besides , they consider how CBIS helps in decision making or solving problems by using CBIS in the decision making process in organizations, which evolved from TPS, MIS, DSS, GDSS, ES, ERP, SCADA and MMS. For the first research question the previous scholars emphasized the importance and necessity of CBIS for decision makers. The researcher is interested to find weather decision makers use CBIS in organizations in Jordan. The preliminary study was done and interviews were conducted in Jordan in October 2009.

III. INTERVIEW PART

The aim of this interview is only to help the researcher to identify the use of CBIS of his research in Jordan, and to test factors for the decision making process of CBIS. A face to face interview was used as a tool to collect preliminary data only. The scope for this interview was limited to decision makers at different levels in the organizations, in using information communication technology in their work in Jordan. Structured interview or what known also as standardized interview is a qualitative approach method, which ensures each interview is done with exactly the same questions in the same order. For this structured interview was considered to be more reliability and validity from the un-structured interviews [28, 29, 30, 31 & 44]. Also, structured interview method was used in a study conducted in five Arab countries [32].

The lack of use of CBIS was observed in many countries in decision making. A study held in Saudi Arabia by [36] confirmed the lack of CBIS use and the need for heavily use for MIS which is one type of CBIS in decision process. Up to the knowledge, no exist for researches done to explore or identify CBIS use for decision makers in organizations in Jordan.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. The Instrument (Interviews).

Face-to-face interviews were conducted, each starting with greeting and enveloped with politeness. An introduction was given about the research for 3-5 minutes. The researcher took notes without biasing the interviewees to any answer and made sure that the time was not too long i.e. each interview lasted between 10-15 minutes and ended with thanking the participants. After one paragraph of the topic title and the researcher name and university, two parts were asked to the interviewees, firstly demographic information, and then followed by four open ended questions; see Appendixes A, B please.

B. Population and Sampling

The researcher tried to do the interview through ten organizations, from the framed population registered ICT organizations which were 170 organizations, after calling the human resources in each organization from the sample, only five of them agreed. Agreement by telephone calling was resulted from five organizations. For non-probability design, it is recognized for two categories: Convenience sampling and purposive sampling and the purposive sampling has two major types: judgment and quota sampling. In this interview a judgment sampling was used [44].

C. Methodology

Face-to-face interviews were conducted, structured interviews as mentioned before have more reliability and validity over the un-structured interviews, and qualitative approach with a judgment type from purposive sampling technique was used for the specific respondents i.e. decision maker using CBIS in organization. Notes were taken by the researcher; this issue was discussed by Sekaran [44] she mentioned:

"The interviews can be recorded in tape if the respondent has no objection. However, taped interviews might bias the respondents` answers because they know their voices are being recorded" (P. 231).

The interview technique was used for each starting with greeting and enveloped with politeness. An introduction was given about the research for 3-5 minutes. The researcher took notes without biasing the interviewees; each interview lasted between 10-15 minutes and ended with thanking the participants.

Translation process was after confirming the questions from specialist from the Computing School from UUM University as follows:

- An academic translation center in Irbid - City in north part of Jordan from English to Arabic and checked for understandability of meaning.
- Translation was then made from Arabic to English and was compared for possible differences.
- Finally, the corrections needed were made to have the final version in Arabic to insure the reliability and validity [33, 34 & 35].

D. Data collection and Analysis

Despite the richness information that can be collected from qualitative methods, there are some issues and problems to deal with qualitative data [45]. Gathering (association) same answered-questions, after that tabulating data in table was made [42, 44], data was grouped and tabulated to make a sense. A simple descriptive analysis was made for the frequencies of the participants' answers. For the demographic and actual use it is good to be analyzed within descriptive analysis. Whereas, rest of the questions, was good to look out for them nearly from the point of views of Morgan and Smircich in [46] as ontologies or epistemologies i.e. keywords in the beginning of papers or common frequent words in content analysis after tabulating the same answers.

E. Findings

1) *Demographic information:*
From 18 respondents only 2 were females with (11%) and 16 males with (89%), the youngest respondents age was 29, while the eldest age was 55 with Age-Average age 39.8 years for the respondents. The respondents managerial levels was 8 low-level with (33%) and 9 middle-levels with (50%), while, only 3(17%) only were from top-levels.

2) *Computer-based information system Use:*
From 18 participants only 6 with (33.3%) of them declared they are using the CBIS in processing their decisions in their organizations, which means 12 with (66.7%) of the managers are not using CBIS in decision processing in those five organizations.

3) *Advantages of CBIS:*
For the third question, the answers of the CBIS-Users (managers), they mentioned the following words: "Easily, help, fast, useful, and integrated". While, for the managers who did not use CBIS, they mentioned words as: "no need, do not know about, think will be good in future, and good to use future".

4) *Decision making factors:*
The associated answers words for this question were "time, reduce cost, risk, benefits, and resource", and less appearance for "rules, customer, and data".

5) *Softwares and tools of CBIS:*
For the managers who are using CBIS the appearance was for "Spreadsheets, dashboard, business object, integrated system, oracle, and service oriented architecture".

A summary of the demographic information and the answers for the use part are categorized in the following table. 3. It is important to mention here that the interviews were in Arabic and what is mentioned in English the language of publication. In addition, based on Talji [43] the findings were categorized.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE 3. DEMOGRAPHIC INFORMATION AND CBIS USE.

Participants of organizations	Gender	Age	Managerial Level	CBIS Use
Participant 1	male	34	Middle	Yes
Participant 2	male	40	Middle	No
Participant 3	female	39	Low	No
Participant 4	male	33	Low	No
Participant 5	male	45	Middle	Yes
Participant 6	male	46	Top	Yes
Participant 7	male	43	Low	No
Participant 8	male	45	Middle	N0
Participant 9	Male	32	Low	Yes
Participant 10	Male	37	Middle	No
Participant 11	Male	36	Low	No
Participant 12	Male	29	Low	Yes
Participant 13	Male	55	Top	N0
Participant 14	Female	34	Low	N0
Participant 15	Male	39	Middle	Yes
Participant 16	Male	41	Low	N0
Participant 17	Male	46	Top	N0
Participant 18	Male	41	Middle	N0

F. Results and Discussion

The purpose of these interviews was to identify the Use of CBIS in decision making in organizations in Jordan, and to test some factors in a proposed model. The researcher ensured that all the participants are decision makers (managers) at any level, and that, all the randomly selected organizations are inclined towards information and communication technology (ICT) i.e. they are using the facility of the technology or have the lowest level of technology. For example, the organization has a website, or uses the internet, and /or the employees have Pc`s in their workplace.

Decision making factors as: time, cost, risk, benefits, and resources are wanted in any try to introduce model for the decision makers, these factors were review by Ashakkah and Rozaini [37]. In addition, the appearance of these factors was recognized with the decision makers whom are users of CBIS answers. CBIS is encouraged to be adopted and used for its benefits as cutting cost, saving time, and making the work easier. And for the tools of CBIS, spreadsheets appeared as a low level while dashboard was for top levels of decision makers. Returning back to the aim of this paper, the CBIS adoption and use need future researches to explore its roles for decision makers, up to the knowledge of the researcher no previous reaches was done in the CBIS in decision making in organization in Jordan. Whereas, for the ICT area asserted that ICT in Jordan need more interest, in order to develop country like Jordan, there is an increasing need to give more interest in ICT development area [38]. Which implies the CBIS use for the decision makers in Jordan interest also is needed, since the CBIS need ICT infrastructure availability as a basic root in organizations.

IV. CONCLUSION AND FUTURE RESEARCH

From the Interviews conducted in five organizations in Jordan with the decision makers (managers) in different managerial levels, the aim was to collect a Preliminary data to find issues about CBIS in decision making in organizations in Jordan, and to help the researcher to test some factors in the proposed model. The researcher conducted 18 face-to face interviews in five ICT organizations through which he was keen not to be biased with the participants in any answer. All along, the participants were assured that their answers would only be used for the research purposes, including names of people and organizations that were promised not to be declared. Lastly, many factors were found to affect the CBIS in decision making from the results of the 18 interviewees, only 6 of them were using the CBIS. Which mean the adoption and use of the CBIS system in decision making in Jordanian organizations still needs more focus and further research.

These interviews have some limitations as the sample size and the self reporting, in all, other view by Delone and Mclean [40, 41] for the updated IS success model, it was a revised for the "Use" to be "intention to use and use" and to put the "benefits" as an output, so it is good to adapt a technology theory which involves the Use and Intention to Use in a future research model, this open the door for researchers to do more researches with this view.

ACKNOWLEDGMENT

The authors wish to acknowledge the reviewers in IJCSIS technical committee for valuable comments, and thank them for their efforts.

REFERENCES

[1] R. Stair and G. Reynolds, "*Principles of Information Systems*," 7th ed. Boston, MA: Thomson Course Technology, 2006.

[2] R. Stair and G. Reynolds, "*Principles of Information Systems*," 9th ed. Boston, MA: Thomson Course Technology, 2010.

[3] G. Mentzas, "A functional taxonomy of computer based information systems" ,*International Journal of Information Management.* Vol. 14, No. 6, PP. 397-410, 1994.

[4] F. Mahar, "Role of information technology in transaction processing system", *in Pakistan Journal of Information and Technology*, Vol. 2, No. 2, PP. 128-134, 2003.

[5] F. Fuller and W. Manning, " *Computer and Information Processing*", USA: International Thomson Publishing, 1994.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[6] T. J. Murray, " *Computer Based Information Systems*," USA: HomeWood, Illinios, 1985. P. 12.

[7] D. W. Walker, "*Computer Based Information Systems: An introduction*," Australia: Pergamon Press , 1988.

[8] R. R. Nelson and H. P. Cheney, "Educating the CBIS User : A Case Analysis," *Data Base*, Vol. 18, No. 2, PP. 11-16 . Winter 1987.

[9] H. B. Eom, S. M. Lee and E-H. Suh, " Group decision support systems: An essential tool for resolving organizational conflicts," *International Journal of Information Management*, Vol. 10, No. 3, PP. 215-227, *1990*.

[10] G. E. Vlahos and T.W. Ferratt, "Information technology use by managers in Greece to support decision making: Amount, perceived value, and satisfaction," *Information & Management*, Vol. 29, No. 6, PP. 305-315. 1995.

[11] C. Goodwin, " The impact of a computer based information system (CBIS) on foreign investments opportunities," in *Proceedings of 12th Annual Conference of the International Academy for Information Management*. USA, PP. 362-367, 1997.

[12] J. Wong and T. Du, " Project-centered teaching on CBIS to IBBA students in Hong Kong," *SIGCSE Bulletin*.Vol. 35, No. 4, PP. 35-38, 2003.

[13] G. E. Vlahos, T. W. Ferratt and G. Knoepfle, "The use of computer-based information systems by German managers to support decision making," *Information & Management*, Vol. 41, No. 6, PP. 763-779, 2004.

[14] K. C. Laudon and J. P. Laudon, " *Essential of Management Information Systems: Managing the Digital Firm,"* sixth ed. New Jersey: Prentice-Hall, Pearson, 2005.

[15] G. Dhillon, "Gaining benefits from IS/IT implementation: Interpretations from case studies", *International Journal of Information Management*, Vol. 25, PP. 502-515, 2005.

[16] E. Turban, J. Aronson, T. Liang and R. Sharda, " *Decision Support and Business Intelligence Systems*," 8th ed. New Jersey: Prentice-Hall, Pearson, 2007.

[17] E. Turban, J. Aronson, T. Liang and R. Sharda, " *Decision Support and Business Intelligence Systems*," 9th ed. New Jersey: Prentice-Hall, Pearson, 2011.

[18] M. G. Martinsons and R. M. Davison, "Stratigic decision support systems: comparing American, Japanese and Chinese management," *Decision Support Systems*, Vol. 43, No. 1, PP. 284-300, 2007.

[19] T. Liang, " Recommendation systems for decision support: An editorial introduction," *Decision Support Systems*, Vol. 45, PP. 385-386, 2008.

[20] S. Persson, L. Malmgren and H. Johnsson, "Information management in industial housing design and manufacture," *Journal of Information Technology in Construction (ITcon)*, Vol. 14, PP. 110-122, 2009.

[21] N. Dlodlo, L. Hunter, C. Cele, F. A. Botha and R. Metelerkamp, "A decision support system for wool classification," *AUTEX Research Journal*, Vol. 9, No . 2, PP. 42-46, 2009.

[22] S. Liu, B. H. A. Duffy, I. R. Whitfield and M. I. Boyle, "Integration of decision support systems to improve decision support performance," in *Knowledge and Information Systems*, Springer London. Vol. 22, No. 3, PP. 261-286, 2009.

[23] J. G. Fogarty and B. Armstrong, "Modeling the interactions among factors that influence successful computerisation of small businesses," *Australasian Journal of Information Systems*, Vol. 15, No. 2, pp. 73-89, 2009.

[24] A. S. Carlsson, S. Hrastinski, S. Henningsson and C. Keller, "An approach for designing management support system: the design science research process and its outcomes," *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*, 2009.

[25] M. M. N. Al-ahmad Malkawi, N. M. Alraja and T. Alkhayer, " Information systems auditing applied study at banks listed in the Damascus stock exchange Syria," *European Journal of Economics, Finance and Administrative Sciences*, No. 21, pp. 119-132, 2010.

[26] H-S. Kim, Y-G. Kim and C-W. Park, "Integration of firm`s resource and capability to implement enterprise CRM: A case study of a retail bank in Korea," *Decision Support Systems*, Vol. 48, No. 2, PP. 313-322, 2010.

[27] S. Patel, J. Zaver, "A risk assessment model for cyber attack on information system," *Journal of Computers*, Vol. 5, No. 3, pp. 352–359, 2010.

[28] M. A. Campion, E. P. Pursell and B. K. Brown, "Structured interview : Raising the psychomeric properities of the employment interview," Personnel Psychology, Vol. 41, PP. 25-42, 1988.

[29] M. A. Campion, D. K. Palmer and J. E. Campion, "A review of structure in the selection interview : Raising the psychomeric properities of the employment interview," Personnel Psychology, Vol. 50, PP. 655-702, 1997.

[30] P. E. Lowry, "The structured interview: An alternative to the assessment center?," *in Public Personnel Management*, Vol. 23, No. 2, PP. 201-215, 1994

[31] M. A. Campion, D. K. Palmer and J. E. Campion , "A review of structure in the selection interview," *Personnel Psychology*, Vol. 50, PP. 655-702, 1997.

[32] C. E. Hill, K. D. Loch, D. W. Straub and K. El-Sheshai, " A qualitative assessment of Arab culture and information technology transfer," *Journal of Global Information Management*, Vol. 6, No. 3, PP. 29–38, 1998.

[33] R. Brislin, "Comparative research methodology: Cross-cultural studies," *International Journal of Psychology*, vol. 11, no. 3, pp. 215-229, 1976.

[34] E. Abu-Shanab and M. Pearson**,** "Internet banking in Jordan: An arabic instrument validation process," *The International Arab Journal of Information Technology*, Vol. 6, No. 3, PP. 235-246, July 2009.

[35] E-S. Cha, K. H. Kim, and J. A. Erlen, " Translation of scales in cross-cultural research: Issues and techniques," *Journal of Advanced Nursing*, Vol. 58, No. 4,PP. 386–395, 2007.

[36] S. Al-Zhrani, " Management information systems role in decision-making during crises: case study," *Journal of Computer Science*, Vol. 6, No. 11, PP. 1247-1251, 2010.

[37] M. S. AL-Shakkah and W. Rozaini, "Empirical study of evolution of decision making factors from 1990-2010," *International Journal of Computer Science and Information Security*, Vol. 9, No. 9, PP. 59-66, 2011, USA.

[38] S. Mofleh, M. Wanous and P. Strachan, " Developing countries and ICT initiatives: Lesson learnt from Jordan's experience. *Electronic Journal of IS in Developing Countries*, Vol. 34, No. 5, PP. 1-17, 2008.

[39] S. Al-Jaghoub and C. Westrup, "Jordan and ICT led development: toward a cometition state," Information Technology and People, Vol. 16, No. 1, PP. 93-110, 2003.

[40] W. H. DeLone and E. R. McLean, "The DeLone and McLean model of information systems success: A Ten-Year update,. *Journal of Management Information Systems*, Vol. 19, No. 4, pp. 9-30, 2003.

[41] W. H. DeLone and E. R. McLean, "Information systems success: The quest for the dependent variable," *Information Systems Research* , Vol. 3, No.1, pp. 60-95, 1992.

[42] R. k. Yin, "The case study crisis: Some answers," *Administrative Science Quarterly*. Vol. 26, No. 1, PP. 58-65, 1981.

[43] S. Talja, " Analyzing qualitative interview data: The discourse analytic method," *Library & Information Science Research*, Vol. 21, no. 4, PP. 459–477, 1999.

[44] U. Sekaran, "*Research Methods for Business: A Skill Building Approach*," (Fourth Ed.), USA: John Wiley & Sons, Inc. 2003.

[45] M. B. Matthew, "Qualitative data as an attractive nuisance: The problem of analysis," *Administrative Science Quarterly*, Vol. 24, No. 4, pp. 590-601, 1979.

[46] G. Morgan and L. Smircich, "The case for qualitative research," *Academy of Management Review*, Vol. 5 , No. 4, PP. 491-500, 1980.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

APPENDIXES

APPENDIX. A Questions for Structured Interview English Version.

Research Interview

Dear Sir/Madam:

This is an Interview for the "Role of Computer-Based Information System (CBIS) in Decision making in your Organizations", for Mohammed Suliman Shakkah, a PhD student from UUM University , Malaysia; firstly we would like to thank you for your participation and your time. Please respond to all of the questions. We are grateful for your cooperation and rest assured that all responses will be only for academic research (No names of persons or organizations will be used).

Q1: demographic information:

1. □ Male □ Female
2. Age your age about……..
3. Managerial level □ Top □ Middle. □ low

Q2: Are you using the CBIS in your decision making process in your organization?

□ Yes □ No

……

……

Q3: What are the advantages of using the CBIS in decision making in your opinion?

……

……..

Q4: In the decision making process, what do you think are the major factors or issues to look for?

……

……

Q5: What is the software that you use in processing your decisions? (If you are a CBIS user).

……

……

Thank You.

APPENDIX B. Questions for Structured Interview Arabic Version

مقابلة لبحث دكتوراه

السيد / السيده المحترم

هذه المقابله لبحث حول دور نظم المعلومات المحوسبه في عمل القرارات في شركتك، لطالب الدكتوراه :محمد سليمان الشقاح من جامعه الشمال الماليزيه بماليزيا ، اولا اود ان اشكركم على مشاركتكم ووقتكم. نود منكم الاجابه على جميع الفقرات ان امكن، نحن ممتنون لتعاونكم.

جميع الاجابات ستستخدم فقط لاغراض البحث العلمي،بمعنى الاسماء للاشخاص والشركات لن تذكرز

س1: معلومات شخصيه:

اولا: الجنس : 1- ذكر 2- أنثى.

ثانيا: كم العمر؟ .....

ثالثا: المستى الاداري: 1- عالي 2- متوسط 3- منخفض

س2 : هل تستخدم نظم المعلومات المحوسبه في عمل القرارات في شركتك؟

1- نعم 2- لا

……

……

س3: ما هي الفوائد من اسخدام نظم المعلومات المحوسبه في عمل القرارات في شركتك برأيك؟

……

……

س4 : عند معالجه عمل القرارات، ما هي العومل او الاشياء التي يجب ان تؤخذ بالحسبان؟

……

……

س5 : ما هي البرامج التي تستخدمها عند عمل القرارات؟(اذا كنت هل تستخدم نظم المعلومات المحوسبه).

……

……

شاكرين لكم تعاونكم.

AUTHORS PROFILE

Mohammed Suliman Al-Shakkah Received the B.Sc degrees in Maths from yarmouk university in 1998, MSc in Information Technology (IT) from Universiti Sins Malaysia (USM) in 2007, he is vice-dean and lecturer from (2009-2011) in Alghad International Colleges for Health and Medical Sciences in Kingdom of Saudi Arabia. He is a candidate PhD student in the final stage, started in 2007 Universiti Utara Malaysia (UUM), interested in decision support system (DSS), decision processing for managers in organizations with structuaral equation modeling (SEM) technique, adoption, acceptance and barriers use of computer-based information system (CBIS) in developing countries.

Dr. Wan Rozaini Received the B.Sc degrees in Physics from Universiti Sins Malaysia (USM) in 1982, PG Diploma in Systems Analysis for public Sector from Universiti of Aston in 1983 in UK. She received MSc, ORSA at UK in 1984. PHD, MIS from Universiti of Salge in UK 1996. Now she Associate professor in Universiti Utara Malaysia and Director of ITU- UUM, ASP COE for Rural ICT development

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

An Immune Inspired Multilayer IDS

Mafaz Muhsin Khalil Alanezi
Computer Sciences
College of Computer Sciences and Mathematics
Iraq, Mosul, Mosul University
mafazmhalanezi@gmail.com

Najlaa Badie Aldabagh
Computer Sciences
College of Computer Sciences and Mathematics
Iraq, Mosul, Mosul University
najladabagh@yahoo.com

Abstract—The use of artificial immune systems in intrusion detection is an appealing concept for two reasons. Firstly, the human immune system provides the human body with a high level of protection from invading pathogens, in a robust, self-organized and distributed manner. Secondly, current techniques used in computer security are not able to cope with the dynamic and increasingly complex nature of computer systems and their security.

The objective of our system is to combine several immunological metaphors in order to develop a forbidding IDS. The inspiration come from: (1) Adaptive immunity which is characterized by learning, adaptability, and memory and is broadly divided into two branches: humoral and cellular immunity. And (2) The analogy of the human immune systems multilevel defense could be extended further to the intrusion detection system itself. This is also the objective of intrusion detection which need multiple detection mechanisms to obtain a very high detection rate with a very low false alarm rate.

Keywords: Artificial Immune System (AIS); Clonal Selection Algorithm (CLONA); Immune Complement Algorithm (ICA); Negative Selection (NS); Positive Selection (PS); NSl-KDD dataset.

I. INTRODUCTION

When designing an intrusion detection system it is desirable to have an adaptive system. The system should be able to recognize attacks it has not seen before and then respond appropriately. This kind of adaptive approach is used in anomaly detection, although where the adaptive immune system is specific in its defense, anomaly detection is non-specific. Anomaly detection identifies behavior that differs from "normal" but is unable to the specific type of behavior, or the specific attack. However, the adaptive nature of the adaptive immune system and its memory capabilities make it a useful inspiration for an intrusion detection system [1].

However on subsequent exposure to the same pathogen, memory cells are already present and are ready to be activated and defend the body. It is important for an intrusion detection system to be adaptive. There are always new attacks being generated and so an IDS should be able to recognize these attacks. It should also then be able to use the information gathered through the recognition process so that it can quickly identify the attacks in the future [1].

Dasgupta et. al. [2, 3] in which they describe the use of several types of detector analogous to T helper cells, T suppressor cells, B cells and antigen presenting cells in two type of data binary and real, to detect anomaly in time series data generated by Mackey-Glass equation.

NSL-KDD are data Sets provide platform for the purpose of testing intrusion detection systems and to generate both background traffic and intrusions with provisions for multiple interleaved streams of activity [4]. These provide a (more or less) repeatable environment in which real-time tests of an intrusion detection system can be performed. The data set contain records each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type, The data set contains 24 attack types. These attacks fall into four main categories: DoS; U2R; R2L; and Probing [24, 26]. These data set available at [25].

II. IMMUNITY IDS OVERVIEW

In computer security there is no single component or application that can be employed to keep a computer system completely secure. For this reason it is recommended that a multilevel defense approach be taken to computer security. The biological immune system employs a multilevel defense against invaders through nonspecific (innate) and specific (adaptive) immunity. The problems for intrusion detection also need multiple detection mechanisms to obtain a very high detection rate with a very low false alarm rate.

The objective of our system is to combine several immunological metaphors in order to develop a forbidding IDS. The inspiration come from: (1) Adaptive immunity which is characterized by learning, adaptability, and memory and is broadly divided into two branches: humoral and cellular immunity. And (2) The analogy of the human immune systems multilevel defense could be extended further to the intrusion detection system itself.

An IDS is designed with three phases: Initialization and Preprocessing phase, Training phase, Testing phase. But the Training phase has two defense layers, the first layer is a Cellular immunity (T & B cells reproduction) where an ALCs would attempt to identify the attack. If this level was unable to identify the attack the second layer Humoral immunity (Complement System), which is a more complex level of detection within the IDS would be enabled. The *complement system*, represents a chief component of innate immunity, not

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

only participates in inflammation but also acts to enhance the adaptive immune response [23]. All memory ALCs obtained from Training phase layers used in Testing phase to detect attacks. This multilevel approach could provide more specific levels of defense and response to attacks or intrusions.

The problem with anomaly detection systems is that often normal activity is classified as intrusive activity and so the system is continuously raising alarms. The co-operation and co-stimulation between cells in the immune system ensures that an immune response is not initiated unnecessarily, thus providing some regulation to the immune response. Implementing an error-checking process provided by co-operation between two levels of detectors could reduce the level of false positive alerts in an intrusion detection system.

The algorithm works on similar principles, generating detectors, and eliminating the ones that detect self, so that the remaining detectors can detect any non-self.

The initial exposure to Ag that stimulates an adaptive immune response is handled by a small number of low-affinity lymphocytes. This process is called primary response and this what will happened in Training phase. Memory cells with high affinity for the encounter, however, are produced as a result of response in the process of proliferation, somatic hyper mutation, and selection. So, a second encounter with the same antigen induces a heightened state of immune response due to the presence of memory cells associated with the first infection. This process is called secondary response and this what will happened in Testing phase. By comparison with the primary response, the secondary response is characterized by a shorter lag phase and a lower dose of antigen required for causing the response, and that could be notice in the run speed of these two phases.

The overall diagram of Immunity-Inspired IDS in figure (1) Note the terms ALCs and detectors have the same meaning in this system.

A. Initialization and Preprocessing phase

Have the following operations:

1) Preprocessing NSL dataset

The data are partitioned in to two classes: normal and attack, where the attack is the collection of all 22 different attacks belonging to the four classes described in section I, the labels of each data instance in the original data set are replaced by either `normal' for normal connections or `anomalous' for attacks. Due to the abundance of the 41 features, it is necessary to reduce the dimensionality of the data set, to discard the irrelevant attributes. Therefore, information gains of each attribute are calculated and the attributes with low information gains are removed from the data set. The information gain of an attribute indicates the statistical relevance of this attribute regarding the classification [21].

Based on the entropy of a feature, information gain measures the relevance of a given feature, in other words its role in determining the class label. If the feature is relevant, in other words highly useful for an accurate determination, calculated entropies will be close to 0 and the information gain will be close to 1. Since information gain is calculated for discrete features, continuous features are discretized with the emphasis of providing sufficient discrete values for detection [20].

The most 10 significant features the system obtained are: duration, src_bytes, dst_bytes, hot, num_compromised, num_root, count, srv_count, dst_host_count, dst_host_srv_count.

a) Information Gain

Let S be a set of training set samples with their corresponding labels. Suppose there are m classes (here m=2) and the training set contains s_i samples of class I and s is the total number of samples in the training set. Expected information needed to classify a given sample is calculated by [20, 21]:

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i}^{m} \frac{s_i}{s} \log_2 (\frac{s_i}{s}) \quad (1)$$

A feature F with values $\{ f_1, f_2, \ldots, f_v \}$ can divide the training set into v subsets $\{ S_1, S_2, \ldots, S_v \}$ where S_j is the subset which has the value f_j for feature F. Furthermore let S_j contain s_{ij} samples of class i. Entropy of the feature F is

$$E(F) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} \times I(s_{1j}, \ldots, s_{mj}) \quad (2)$$

Information gain for F can be calculated as:

$$Gain(F) = I(s_1, \ldots, s_m) - E(F) \quad (3)$$

b) Univariate discretization process

Discrete values offer several advantages over continuous ones, such as data reduction and simplification. Quality discretization of continuous attributes is an important problem that has effects on speed, accuracy, and understandability of the classification models [22].

Discretization can be univariate or multivariate. Univariate discretization quantifies one continuous feature at a time while multivariate discretization simultaneously considers multiple features. We mainly consider univariate (typical) discretization in this paper. A typical discretization process broadly consists of four steps [22]:

- Sort the values of the attribute to be discretized.
- Determine a cut-point for splitting or adjacent intervals for merging.
- Split or merge intervals of continuous values, according to some criterion.
- Stop at some point.

Since information gain is calculated for discrete features, continuous features should be discretized [20, 22]. To this end, continuous features are partitioned into equalsized partitions by utilizing equal frequency intervals. In equal frequency intervals method, the feature space is partitioned into arbitrary number of partitions where each partition contains the same number of data points. That is to say, the range of each partition is adjusted to contain N dataset instances. If a value occurs more than N times in a feature space, it is assigned a

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

partition of its own. In "21% NSL" dataset, certain classes such as denial of service attacks and normal connections occur in the magnitude of thousands whereas other classes such as R2L and U2R attacks occur in the magnitude of tens or hundreds. Therefore, to provide sufficient resolution for the minor classes N is set to 10 [20]. The result of this step are the most gain indexes to use them later in preprocessing training and testing files.

2) Self and NonSelf Antigens

As mentioned in chapter 2 that each record of NSL or KDD dataset contains 41 features and is labeled as either normal or an attack, so it would be here as Self and NonSelf respectively.

The dataset used in the training phase of the system contain about 200 records from normal and attack records, the attack records have records from all types of attack in the original dataset. And this rule applied on NSL and KDD datasets. But the all "21% NSL" test datasets used when test the system in testing phase.

The system in training and testing phase, apply on each file before enter to it: selecting the most gain indexes and convert each continuous feature to discrete.

3) Antigens Presentation

T cell and B cell are assumed that recognize antigens in different ways. In biological immune system, T cells can only recognize internal features (peptides) processed from foreign protein. In our system, T cells recognition is defined as bit-level recognition (real, integer). This is a low-level recognition scheme. In the immune system, however, B cells can only recognize surface features of antigens. Because of the large size and complexity of most antigens, only parts of the antigen, discrete sites called epitopes, get bound to B cells. B-cell recognition is proposed that is a higher-level recognition (string) at different non-contiguous (occasionally contiguous) positions of antigen strings.

So different data types are used for each ALC in order to compose several detection levels. In order to present the self and nonself antigens on ALCs, there are also converted to suit different data types of ALCs, like integer for T-helper cells, string for B-cells, and real [0-1] for T-suppresser cells .

Real values would be in range [0-1], so Normalization is used for conversion operation.

4) Normalization

Data transformation such as normalization may improve the accuracy and efficiency of classification algorithms involving neural networks, mining algorithm, or distance measurements such as nearest neighbor classification and clustering. Such methods provide better results if data to be analyzed has been normalized, that scaled to specific ranges such as (0-1) [8, 9]. If using the neural network back propagation algorithm for classification mining, normalizing the input values for each attribute measured in the training samples will help speed up the learning phase. For distanced-based methods, normalization helps prevent attributes with initially large ranges from outweighing attributes with initially smaller ranges [9]. There are many methods for data normalization include min-max normalization, z-score normalization, Logarithmic normalization and normalization by decimal scaling [8, 9].

Min-max normalization: The Min-max normalization performs a linear transformation on the original data. Suppose that min_a and max_a are the minimum and the maximum values for feature A. Min-max normalization maps a value v of A to v' in the range [new-min_a, new-max_a] by computing [9]:

$$v' = ((v - min_a) / (max_a - min_a)) * (new\text{-}max_a - new\text{-}min_a) + new\text{-}min_a \quad (4)$$

In the case range is [0-1] the equation would be:

$$v' = (v - min_a) / (max_a - min_a) \quad (5)$$

In order to generalization all the comparisons (NS & PS) done in IIDS, and to simplify the chosen of thresholds values, the calculated affinities between each one of ALCs and all Ags is normalized into the range [1-100] in case Th and B cells, and normalized into the range [0-1] in case Ts cells and CDs.

5) Detector Generation Mechanism

All Nonself or attack records in training file will be consider as the initial detectors (or ALCs) then in training phase eliminates those that match self samples.

Sure there are three types of detectors (integer, string, real). The output of this step is a specified number for every detectors types and their length equal to Self and NonSelf patterns length's which is the number of gain indexes.

6) Affinity Measure by Matching Rules

In several next steps affinity needs to be calculated the between (ALCs & Self patterns) and (ALCs & NonSelf Ags), so matching rules are determined depend on the data type.

- The affinity between an Th ALC (integer) and a NonSelf Ags or Self patterns is measured by *Landscape-affinity matching (Physical matching rule*) [11, 12, 10]. The Physical matching gives an indication of the *similarity* between two patterns, i.e. a higher affinity value between an ALC and a NonSelf Ags implies a stronger affinity.

$$f_{\text{physical}} = \sum_{i=1}^{N} (X_i - Y_i) + 3 \times |\mu|, \quad (6)$$
$$\text{where } \mu = \min(\forall i, (X_i - Y_i)).$$

- The affinity between an Ts ALC (real) and a NonSelf Ags or Self patterns is measured by *Euclidean distance* [11 ,13, 12]. The Euclidean distance gives an indication of the *difference* between two patterns, i.e. a lower affinity value between an ALC and a NonSelf Ags implies a stronger affinity

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2} = \|x - y\| \quad (7)$$

- The affinity between an B ALC (string) and a NonSelf Ags or Self patterns is measured by *R-Contiguous* string matching rule. If *x* and *y* are equal-length strings defined

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

over a finite alphabet, *match(x, y)* is true if *x* and *y* agree in at least *r* contiguous locations [11, 14, 12, 15]. The R-Contiguous string matching gives an indication of the *similarity* between two patterns, i.e. a higher affinity value between an ALC and a NonSelf Ags implies a stronger affinity.

B. Training Phase

Here the system will be train by a serious of recognition operations between the previous generated detectors and self and nonself Ags to constitute multilevel recognition, make the recognition system more robust and ensures efficient detection.

1) First Layer-Cellular immunity (T & B cells reproduction)

Both B cells and T cells undergo proliferation and selection and exhibit immunological memory once they have recognized and responded to an Ag. All system's ALCs progress in the following stages:

a) Clonal and Expansion

Clonal selection in AIS is the selection of a set of ALCs with the highest calculated affinity with a NonSelf pattern. The selected ALCs are then cloned and mutated in an attempt to have a higher binding affinity with the presented NonSelf pattern. The mutated clones compete with the existing set of ALCs, based on the calculated affinity between the mutated clones and the NonSelf pattern, for survival to be exposed to the next NonSelf pattern.

- *Selection Mechanism*

The selection of cells for cloning in the immune system is proportional to their affinities with the selective antigens. Thus implementing an affinity proportionate selection can be performed probabilistically using algorithms like the *roulette wheel selection,* or other evolutionary selection mechanism can be used, such as *elitist selection, rank- based selection, bi-classist selection, and tournament selection* [5].

Here the system use elitist selection because it needs to remember good detectors and discard bad ones if it is to make progress towards the optimum. A very simple selector would be to select the top N detectors from each population for progression to the next population. This would work up to a point, but any detectors which have very high affinity will always make it through to the next population. This concept is known as elitism.

To apply this idea four selected percent values are specified, which determine the percent from each type of ALCs will be select to Clonal and Expansion operations,

$$SelectedALCNo = (ALC_{size} * selectALC_{percent}) / Maxgeneration, \quad (8)$$

Where *SelectedALCNo* is no of ALCs will be Selected to clone them, ALC_{size} is the number of ALCs survived from NS and PS in initialization and Generation phase, $selectALC_{percent}$ is a selected percent value it range [10-100%], and *Maxgeneration* is the maximum no of generation used in random generation of ALCs in initialization and Generation phase.

- *Sorting Affinity*

The affinity is measured here between all cloned ALCs and NonSelf Ags. And sort all ALCs in descending order depend on their affinity with NonSelf Ags.

- *Clonal Operator*

Now is a time to clone the previous selected ALCs in order to expand the number of ALCs in training phase, and ALC how has the higher affinity with NonSelf Ags will has the higher Clonal Rate.

Here the clonal rate is calculated for each one of the selected ALCs,

$$TotalCloneALC = \Sigma^{n}_{i=1} ClonalRateALC_i , \quad (9)$$

where

$ClonalRateALC_i = Round (Kscale / i),$ or
$ClonalRateALC_i = Round (Kscale \times i),$ [16]

The choice between the two equation of $ClonalRateALC_i$ depend on how much clones required? *Kscale* is the clonal rate, *Round*() is the operator that rounds the value in parentheses toward its closet integer value, and *TotalCloneALC* is the total no of clones cells.

- *Affinity Maturation (Somatic hypermutation)*

After producing clones from the selected ALCs, these clones alter by a simple mutation operator to provide some initial diversity over the ALCs population.

The process of affinity maturation plays an important role in adaptive immune response. From the viewpoint of evolution, a remarkable characteristic of the affinity maturation process is its controlled nature. That is to say the hypermutation rate to be applied to every immune cell receptor is proportional to its antigenic affinity. By computationally simulating this process, one can produce powerful algorithms that perform a search akin to local search around each candidate solution. In account to this important aspect of the mutation in the immune system: it is inversely proportional to the antigenic affinity [5]. Without mutation the system is only capable of manipulating the ALCs material that was present in initial population [6].

In case Th, and B ALCs, the system calculate mutation rate for each ALCs depend on its affinity with NonSelf Ags, where higher affinity (similarity) has lower mutation rate.

In Ts case, one can evaluate the relative affinity of each candidate ALCs by scaling (normalizing) their affinities. The inverse of an exponential function can be used to establish a relationship between the hypermutation rate α(.) and normalized affinity D*, as described in next equation. In some cases it might be interesting to re-scale α to an interval such as [0 – 1] [5].

$$\alpha(D^*) = exp(-\rho D^*) \quad (10)$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

where ρ is a parameter that controls the smoothness of the inverse exponential, and D^* is the normalized affinity, that can be determined by $D^* = D/D_{max}$. inverse mean lower affinity (difference) has higher mutation rate.

Mutators generally are not as complicated, they tend to just choose a random point on the ALCs and perturb this allele (part of Gene) either completely randomly or by some given amount [6].

To control the mutation operator *mutation rate* is calculated as descried up, which is determine number of allele from ALCs will be mutate. The hypermutation operator for each type of shape-space as follows:

- Integer shape-space (Th): when mutation rate of the current Th-ALC high enough, randomly choose the alleles position from ALC, and replace them with a random integer values. Another case use inversive mutation that might occur between one or more pairs of allele.
- String shape-space (B): when mutation rate of the current Th-ALC high enough, randomly choose the alleles position from ALC, here the allele has length equal R string, so may the entire characters of allele change or part of them with another characters.
- Real shape-space (Ts): randomly choose the alleles position from ALC, and a random real number to be added or subtracted to a given allele is generated

$$m` = m + \alpha(D^*) N(0,\sigma) \quad (11)$$

where m is allele, $m`$ its mutated version, $\alpha(D^*)$ is a function that accounts for affinity proportional mutation.

- *Negative Selection*

A number of the NS algorithm features that distinguish it from other intrusion detection approaches. They are as follows [4]:

- No prior knowledge of intrusions is required: this permits the NS algorithm to detect previously unknown intrusions.
- Detection is probabilistic, but tunable: the NS algorithm allows a user to tune an expected detection rate by setting the number of generated detectors, which is appropriate in terms of generation, storage and monitoring costs.
- Detection is inherently distributable: each detector can detect an anomaly independently without communication between detectors.
- Detection is local: each detector can detect any change on small sections of data. This contrasts with the other classical change detection approaches, such as checksum methods, which need an entire data set for detection. In addition, the detection of an individual detector can pinpoint where a change arises.
- The detector set at each site can be unique: this increases the robustness of IDS. When one host is compromised, this does not offer an intruder an easier opportunity to compromise the other hosts. This is because the disclosure of detectors at one site provides no information of detectors at different sites.
- The self set and the detector set are mutually protective: detectors can monitor self data as well as themselves for change.

The negative selection (NS) based AIS for detecting intrusion or viruses was the first successful piece of work using the immunity concept for detecting harmful autonomous agents in the computing environment.

The steps of NS algorithm are applied here,

- Generated three types of ALCs (Th, Ts, B), and present them together with the set of Self (normal record) patterns to NS mechanism.
- For all the ALCs generated, compute the affinity between each one of ALCs and all Self pattern, The choose of matching rule to measure the affinity depend on ALCs data type representation.
- If the ALC did not match with all self patterns depend on threshold comparison will survive to inter the next step, and the ALCs whose match with any Self pattern will be discard. Each type of ALCs have its own threshold value specially for NS.
- Goto to the first step until reach the maximum number of generations of ALCs.

But here NS is done between the three types of mutated ALCs and Self patterns, because may be some ALCs match Self pattern after mutation.

- *Positive Selection*

The mutated ALCs survived from previous Negative selection will be put here to face the NonSelf Ags (attack records) in order to distinguish which detectors can detect them and also because may be some ALCs not match NonSelf Ags after mutation so there is no need to keep them. The steps of PS algorithm are applied here:

- Present the three types of ALCs (Th, Ts, B) that survive from NS together with the set of NonSelf Ags to PS mechanism.
- For all the ALCs, compute the affinity between each one of ALCs and all NonSelf Ags, The choose of matching rule to measure the affinity depend on ALCs data type representation.
- If the ALC match with all Nonself Ags depend on threshold comparison will survive to inter the Training Phase, and the ALCs whose did not match with any NonSelf Ags will be discard. Each type of ALCs have its own threshold value specially for PS.
- Goto to the first step until apply PS on all ALCs.

- *Immune Memory*

Save all survived ALCs from NS and PS in text files, text files for each types of ALCs (Th, Ts, B). Here the system produce memory cells to protect against the reoccurrence of the same antigens. Memory cells enable the immune system's response to previously encountered antigens (known as the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

secondary response), which is known to be more efficient and faster than non-memory cells' response to new antigens. In an individual these cells are long-lived, often lasting for many years or even for the lifetime of it.

2) Second Layer-Humoral immunity (Complement System)

This layer automatically activated when the first layer terminate, and this layer simulate the classical pathway of the complement system, which is activated by a recognition between antigen and antibody (here detectors). The classical pathway is composed of three phases: Identify phase, Activate phase and Membrane attack phase. These phases and all its step called Immune Complement Algorithm(ICA) describe in details in [23].

In this system the complement detectors progress ICA steps with several additional step designed for it purpose, the objective of ICA is the continuo in generation, cleave, and bind the CD individuals until find the optimal CD individuals. The system's ICA summary here in the following four phases:

- *ICA: Initialization phase*

- Get the Nonself as the initial first one population A0 has a fix number of Complements detectors CDs as individuals their data type are real in range [0-1].
- Stopping conditions: if the current population has contained the desire number of optimal detectors (*CDsn*) or achieved the maximum generation, then stop, else, continues.
- Define the following operators
  1. Cleave operator O_C: A CD individual cleave according to a cleaved probability P_c, is cleaved in two sub-individuals: a_1 and a_2.
  2. Bind operator O_B : There are two kinds of bind ways between individuals *a* and *b*:
     - Positive bind operator O_{PB} : A new individual $c = O_{PB}(a,b)$
     - Reverse bind operator O_{RB} : A new individual $c = O_{RB}(b,a)$

- *ICA: Identify Phase*

- *Negative Selection*: For each Complement detector in the current population apply NS with Self patterns, and the Complement detector whose match with any Self pattern will be discard. The Euclidean distance used here, which is give an indication of the *difference* between the two patterns, i.e. if the affinity between one CD and all Self patterns *exceed* a threshold, then the detector survive, else discard.
- *Split Population*: isolate the CDs how survived from NS alone (*A0NS*) from the CDs how discarded (*A0PS*).
- *Positive Selection*: For each Complement detector in the *A0NS* apply PS with NonSelf Ags, and the Complement detector whose match with all NonSelf Ag will be survive. The Euclidean distance used here, which is give an indication of the *difference* between the two patterns, i.e. if the affinity between one CD and all NonSelf Ags *not exceed* a threshold, then the detector successfully detect, else not successfully detect.
- *Immune Memory*: if there are successful CD, then store all CDs can detect NonSelf Ags in PS in text file and go to stopping Condition: have an *CDsno optimal complement detectors*, else continues.
- *Sorting CDs*: according to the affinities calculated in previous PS step, Sort all the successful individuals CDs in *A0NS* by their ascending affinities (the higher affinity is the lower value because this affinity is a difference value).
- Immerge Population: first put *A0NS* in the population and then append *A0PS* after it.

- *ICA: Active phase*

- Divide the Population into A_t^1& A_t^2 using *Div* active variable. $A_t^1$is a Cleave Set, and A_t^2 is a Bind Set.
- For each individual in A_t^1apply a Cleave Operator O_C to produce two sub-individual a_1 and a_2. Then take the second sub-individual a_2 for all CD individuals in $A_t^1$and bind them in one remainder cleave set b_t by *Positive bind operator* O_{PB}.

- *ICA: Membrane attack process*

- Using *Reverse bind operator* O_{RB}, bind b_t and each DC individual of A_t^2 to get a membrane attack complex set C_t.
- For each DC individual of C_t , recode it by the code length of initial DC individual, then gets a new set C'.
- Create a random population of complement individuals D, then join them into C', to finally form a new set $E = C' \cup D$. For the next loop A_0 is replace with E .
- If the iteration step not finish go to stopping condition.

C. Testing Phase

This phase apply test on the immune memory of ALCs created in training phase. So here the meeting between memory ALCs and all types of Antigens Selfs and NonSelfs take place, it is important to note here that memory ALCs not encountered in passed with these new Ags.

The Testing phase use Positive Selection to decide wither an Ag is Selfs or NonSelfs (i.e. normal or attack record) by calculate the affinity between ALCs and the new Ags and compared it with testing thresholds. As in Affinity Measure by Matching Rules section. So if any Ag match any one of ALCs it consider anomaly, i.e. a NonSelf Ags (attack), otherwise it is Self (normal).

Performance Measurement

In learning extremely imbalanced data, the overall classification accuracy is often not an appropriate measure of performance. Metrics are used as *true negative rate*, *true positive rate*, *weighted accuracy*, *G-mean, precision*, *recall*, and *F-measure* to evaluate the performance of learning algorithms on imbalanced data. These metrics have been widely used for comparison and performance evaluation of

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

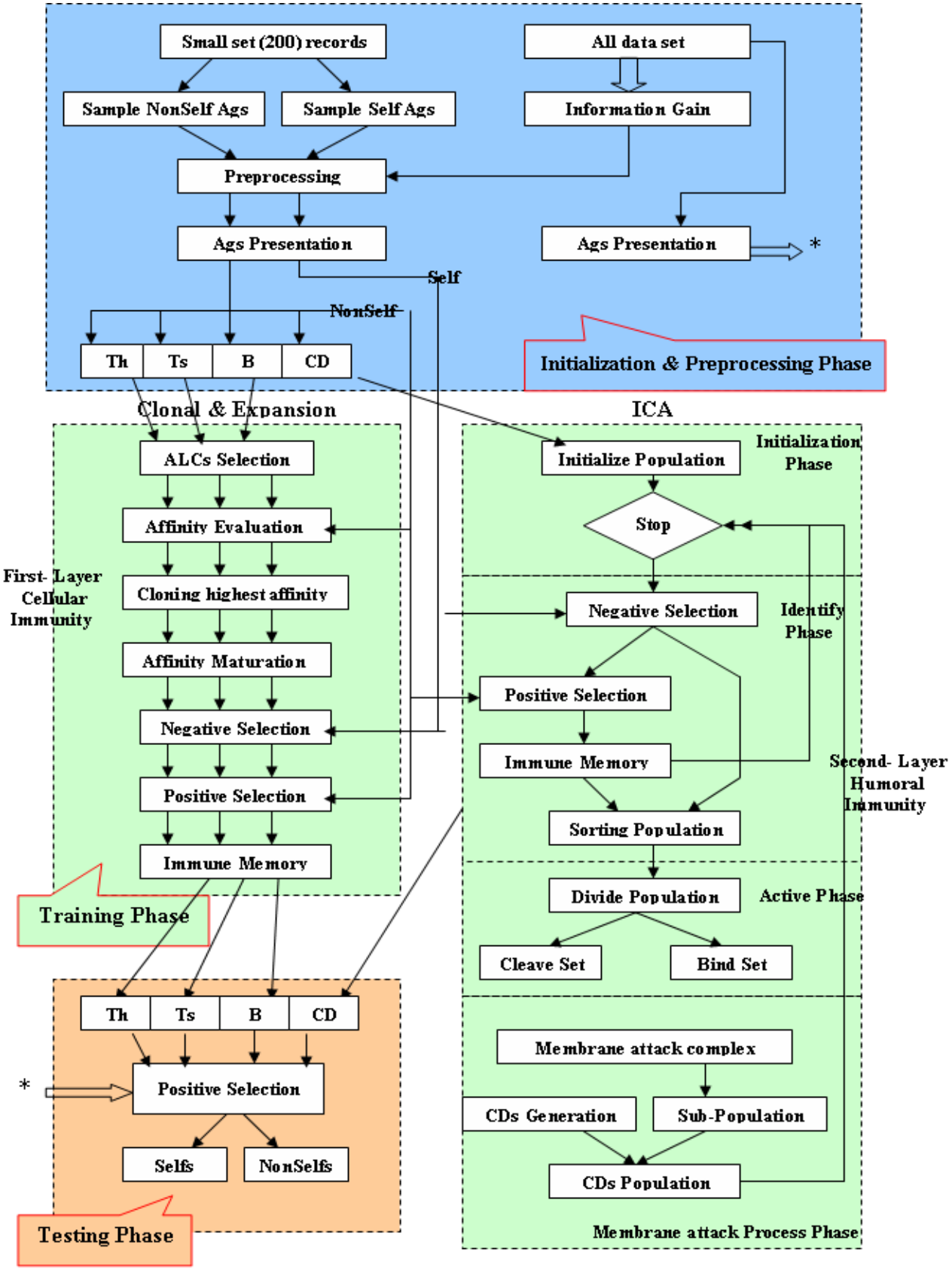


Figure (1): The overall diagram of Immunity IDS.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

classifications. All of them are based on the confusion matrix as shown at table (1) [7, 17, 18, 19].

Table (1): The Confusion matrix.

	predicted positives	predicted negatives
real positives	TP	FN
real negatives	FP	TN

Where TP (true positive), attack records identified as attack; TN (true negative), normal records identified as normal; FP (false positive), normal records identified as attack; FN (false negative), attack records identified as normal [3, 17, 18].

III. IMMUNITY-INSPIRED IDS PSEUDO CODE

Each phase or layer of the algorithm and its iterative processes are given below:

1. Initialization and Preprocessing phase

1.1. Set all parameters that have constant value:

- **Threshold of NS:** Th_{NS} = 60, Ts_{NS} =0.2, Tb_{NS} = 30, $Tcomp_{NS}$ = 0.25;
- **Threshold of PS:** Th_{PS} = 80, Ts_{PS} =0.15, Tb_{PS} = 70, $Tcomp_{PS}$ = 0.15;
- **Threshold of Test PS:** Th_{Test} = 20, Ts_{Test} =0.1, Tb_{Test} = 80, $Tcomp_{Test}$ = 0.05;
- **Generation:** MaxgenerationALC = 500, MaxThsize = 50, MaxTssize = 50, MaxBsize = 25.
- **Clonal & Expansion:** selectTh= 50%, selectTs = 50%, selectB = 100%;
- **Complement System:** MaxgenerationCDs = 1000, PopSize = NonSelfno., CDlength = 10, Div = 70%, CDno = 50;
- **Others:** MaxFeature =10, Interval = 10, classes = 2, ALClength = 10, R-contiguous R = 1, ρ = 2 parameter control the smoothness of exponential (mutation);
- **Classes:**
  - **Normalize class:** contain all functions and operation to perform min-max normalization in range [0-1] and [1-100].
  - **Cleave-Bind Class:** contain Cleave() function O_C ,PositiveBind() function O_{PB}, ReverseBind() function O_{RB}.
- **Input files for Training phase:** NSL or KDD file contain 200 records (60 normal, 140 attack from all attack types).
- **Input files for Testing phase:** files contain 20% from KDD or NSL datasets.

1.2. Preprocessing and Information Gain

- Using the 21%NSL dataset file to calculate the following:
- Split the dataset into two classes normal and attack.
- Convert alphabetic features to numeric.
- Convert all continuous features to discrete, for each class alone.

  For each *one of 41 features* **Do**
  Sort feature's space values;
  Partitioned feature space by *Interval* number specified, each partition contains the same number of data;
  Find the minimum and maximum values;
  Find the initial assignment value
  V = *(maximum-minimum)/Interval no.;*
  Assign each interval i by $V_i = \Sigma_i V$;
  If a value occurs more than *Interval* size in a feature space, it is assigned a partition of its own;
- Calculate Information Gain for every feature in both two classes by applying equations in section 4.3.1.1.
- By selecting the most significant features (*MaxFeature*=10) that have larger values of information gain, the system obtained the same features for both classes (normal and attack) but in different order. So the 10 of the 41 features are continuous and identified as most significant are: 1, 5, 6, 10, 13, 16, 23, 24, 32, 33.
- Save the indexes of these significant feature in text file to use them later in preprocessing the training and testing files.

1.3. Antigens Presentation

- For both training and testing files apply preprocessing operations on the 10 significant features of them.
- Convert all inputted Self & NonSelf Ags to (integer, real, string).
- Apply *Min-Max normalization* on only how has real value to be in range [0-1].

1.4. Detector Generation

- Get NonSelfs Ags as initial Th, Ts, B ALCs, their length is *ALClength* = *MaxFeature*.
- Convert them to 3 type of ALCs (integer, real, string).

2. Training Phase

Input: 200 NSL records (60 normal, 140 attacks from every types);

2.1. First Layer-Cellular immunity (T & B cells reproduction) - Clonal and Expansion

For (all ALCs type**) do**
/*Calculate the select percent for cloning operation;
SelectThNo = (Th_size × SelectTh) / 100;
SelectTsNo = (Ts_size × SelectTs) / 100;
SelectBNo = (B_size × SelectB) / 100;
For (all ALCs type**) do** /* As an example Th
While (*Th_size* < *MaxThsize*) Λ (*generate* < *MaxgenerationALC*)
Calculate the affinity between each ALC and all *NonSelf* Ags;
Sort the ALCs in ascending or descending order (depend on affinity similarity or differently), according to the ALCs affinity;
Select *SelectThNo* of the highest affinity ALCs with all NonSelf Ags as subset *A*;
Calculate Clonal Rate for each one of ALC in *A*, according to the ALCs affinity;
Create clones *C* as the set of clones for each ALC in *A*;
Normalize the *SelectThNo* highest affinity ALCs;
Calculate *mutation Rate* for each one of ALC in *C*, according to the ALCs normalized highest affinity;
Mutate each one of ALC in *C*, according to it's *mutation Rate* and *randomly* select allele no, as the set of mutated clones *C'*;
/*Apply *NS* between mutated ALCs *C'* and Self patterns;
For (*all Self patterns***) do NS**
Calculate affinity by *Landscape-affinity* rule between current Th-ALC & all Self patterns;
Normalize affinities in range [1-100]
If (*all affinity* < Th_{NS})
/* Apply *PS* between survived mutated ALCs from NS and NonSelfs Ags;
For (*all NonSelf Ags***) do PS**
Calculate affinity by *Landscape-affinity* rule between current Th-ALC & all NonSelf Ags;
Normalize affinities in range [1-100]
If (*all affinity* >= Th_{PS})
Th-ALC survive and save it in file "Thmem.txt";
Th_size = Th_size + 1;
Else
Discard current Th-ALC;
Go to next Th-ALC
End If
Add survived mutated ALCs from NS & PS to "Thmem.txt", as *Secondary response*;
generate++;
End While
End For
Call Complement System to activate it;

2.2. Second Layer-Humoral immunity (Complement System)

2.2.A. ICA: Initialization phase

Get NonSelfs as an initial real [0-1] population A_0 has CDs equal *PopSize*.
Stop: if the current population has contained *CDsn* optimal detectors or achieved *MaxgenerationCDs* generation.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Assign a random real value [0.5-1] as Cleave Probability *Pc*;

2.B. ICA: Identify Phase

While ((*CD_size* < *CDsn*) Λ (*generate* <= *MaxgenerationCDs*))

For (*each CD in Population* A_0**) do**

For (*all Self patterns***) do NS**

Calculate affinity by *Euclidean distance* between current CD & all Self patterns;

Normalize affinities in range [0-1]

If (*all affinity* > $Tcomp_{NS}$)

Put current CD in A_0NS sub-population;

Else

Put current CD in A_0Rem sub-population;

End For

For (*each CD in Population* A_0NS**) do**

For (*all NonSelf Ags***) do PS**

Calculate affinity by *Euclidean distance* between current CD & all NonSelf Ags;

If (*all affinity* <= $Tcom_{PS}$)

Save it in file "CDmem.txt";

CD_size = CD_size + 1

Else

Discard current CD;

End For

Sort all CDs in A_0NS by their ascending affinities with *NonSelf Ag*, and put them in *At*;

Append A_0Rem at last *At*;

2.2.C. ICA: Active phase

Divide *At* into A_t^1 and A_t^2 depend on *Div* active variable; /* A_t^1 is a cleave set, A_t^2 is a bind set;

For (*each CD individual in* A_t^1) **do**

Apply cleave operator on CD with cleave probability *Pc* to produce two sub-individual a_1 and a_2, O_C (CD, *Pc*, a_1, a_2);

For (*all sub-individual in* a_2) **do**

Bind them in one remainder cleave set b_t by *Positive bind operator* O_{PB}, $b_t = O_{PB}$ (a_{2i},…,Λ, a_{2n});

2.2.D. ICA: Membrane attack process

For (*each CD individual* a_i *in* A_t^2) **do**

Bind b_t with current individual of A_t^2 by Reverse bind operator O_{RB}, to obtain Membrane Attack complex set C_t, $Ct = O_{RB}(bt, a_i)$;

For (*each individual* c_i *in* C_t) **do**

Recode it to the initial CDlength = 10 to get a new set *C*; /* different strategies may use here for that purpose.

Create Random population of CDs individuals as a set *D*;

Join *C* and *D* in one set *E,* consider it as a new population;

E= C & D,

A0 = E;

Generate++;

End While

3. Testing Phase

Input: 21%NSL dataset;

Initialize**:** *FP, FN, TP, TN, DetectionRate, FalseAlarmRate, ACY, Gmean.*

/*Calculation number of *normalAg* & *attackAg* only for the purpose to calculate performance measurements

For (*each record in input file*) **do**

If (*record type is normal*)

normalAg = normalAg +1;

Else

attackAg = attackAg +1;

/*Antigens Presentation

Convert all inputted Self & NonSelf Ags to (integer, real, string).

Apply *Min-Max normalization* on only how has real value to be in range [0-1].

Read *ThMemory* ALCs;

Read *TsMemory* ALCs;

Read *BMemory* ALCs;

Read *CDMemory* Detectors;

/*Apply PS between all inputted Ags (Self & NonSelf, i.e. normal & attack) and all memory ALCs;

For (*all Thmemory ALCs***) do** /* As an example Th

For (*all Ags types***) do PS**

Calculate the affinity by *Landscape-affinity* rule between each one of Ags and current *Thmemory ALCs*;

Normalize affinities in range [1-100]

If (*affinity* > Th_{NS})

Thmemory ALCs detect a NonSelf Ag;

Record Ag name;

TP = TP + 1; /* no of detected Ags

Else

FP = FP +1;

/*do the previous on, *TsMemory*, *BMemory*, and *CDMemory.*

3.1. Performance Measurement

TN = normalAg - FP;
FN = attackAg – TP;
DetectionRate = TP / (TP + FN);
FalseAlarmRate = FP / (TN + FP);
ACY = (TP + TN)/(TP + TN + FP + FN);
Gmean = DetectionRate × (1 - FalseAlarmRate);
Precision =TP / (TP + FP);
Recall = TP / (TP + FN);
*F-measure = (2 * Precision * Recall) / (Precision + Recall);*

IV. SYSTEM PROPERTIES

The special properties of Immunity IDS are:

- The small size of training data, about 200 NSL records(60 normal, 140 attack from different types).
- The speed of system, where the training periods are about 1 minute because the small size of training data, and the testing periods are about very few minutes depend on memory ALCs size.
- The results of the system test different after each training operation, because it depend on randomly mutation for ALCs.
- The numbers of memory ALCs depend on number of times of retraining, or what the system want.
- The system permit to delete all memory contents to start new training, or every new training after the first one, the ALCs result from it will be add to memory with the previous.
- The detection rate is high with small numbers of memory ALCs produced from one training.
- To apply the Immunity IDS in real, the optimal result of one or more training are chosen, to carry out optimal outcome.
- The thresholds values determined by many experiments until found the fit values.
- The IIDS implemented using C# language.

V. Experimentals Results

1) Several series of experiments were performed by 175 detectors (memory ALCs) sizes. The table (2) shows the test results of 10 training operation done seriously on 200 records to test "NSLTest-21.txt" file, which contain 9698 attack records and 2152 normal records.

2) Comparison of performances (ACY) between single level detection and multilevel detection. The ACY is chosen because it include both TPR and TNR. The table (3) and figure (2) show the test results of 5 training operation done seriously also on "NSLTest%.txt" file. Notice that CDs have the higher

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

accuracy and B cells has the lower accuracy. Although the accuracy of IIDS lower than CD but IIDS has the higher detection rate this return to the effect of false alarm.

Table (2): Results of Test experiments.

TP	TN	FP	FN	TPR	TNR	ACY	g_m.	Prec.	F-m.
8748	2108	44	950	0.9	0.02	0.92	0.88	0.99	0.94
8893	1871	281	805	0.92	0.13	0.91	0.80	0.97	0.94
8748	2123	29	950	0.9	0.01	0.92	0.89	1	0.95
8730	2146	6	968	0.9	0	0.92	0.9	1	0.95
8800	1971	181	898	0.91	0.08	0.91	0.84	0.98	0.94
8788	2014	138	910	0.91	0.06	0.91	0.85	0.98	0.94
8802	2007	145	896	0.91	0.07	0.91	0.85	0.98	0.94
8817	2046	106	881	0.91	0.05	0.92	0.86	0.99	0.95
8833	2002	150	865	0.91	0.07	0.91	0.85	0.98	0.94
8869	1963	189	829	0.91	0.09	0.91	0.83	0.98	0.94

Table 3: Accuracy of IIDS and each type of ALCs.

ACY				
IIDS	Th	Ts	B	CD
0.91	0.84	0.73	0.22	0.92
0.91	0.84	0.78	0.21	0.92
0.91	0.84	0.74	0.22	0.92
0.91	0.84	0.74	0.25	0.92
0.91	0.84	0.77	0.30	0.92

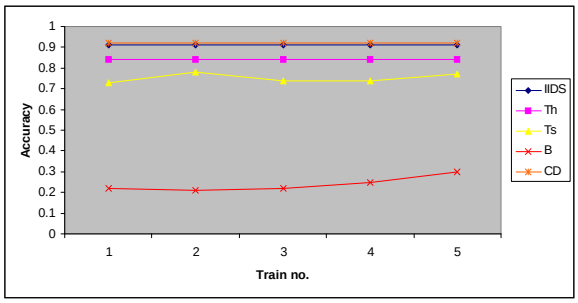


Figure 2 : Accuracy curve comparing the single-level detection (Th, Ts, B, CD) and multilevel (IIDS).

REFERENCES

[1] M. Middlemiss, "*Framework for Intrusion Detection Inspired by the Immune System"*, The Information Science Discussion Paper Series, July 2005.

[2] Dasgupta, D., Yu, S., Majumdar, N.S. Majumdar, "*MILA - multilevel immune learning algorithm*", In Cantu-Paz, E., et. al., eds.: Genetic and Evolutionary Computation Conference, Chicago, USA, Springer-Verlag (2003) 183–194

[3] Dasgupta, D., Yu, S., Majumdar, N.S. Majumdar, "*MILA – multilevel immune learning algorithm and its application to anomaly detection*", DOI 10.1007/s00500-003-0342-7, Springer-Verlag 2003.

[4] Jungwon Kim, Peter J. Bentley, Uwe Aickelin, Julie Greensmith, Gianni Tedesco, Jamie Twycross, "*Immune System Approaches to Intrusion Detection - A Review*", Editorial Manager(tm) for Natural Computing, 2006.

[5] L. N. de Castro and J. Timmis. "*Artificial Immune Systems: A New Computational Intelligence Approach*", book, Springer, 2002.

[6] Edward Keedwell and Ajit Narayanan, "*Intelligent Bioinformatics The application of artificial intelligence techniques to bioinformatics problems*", book, John Wiley & Sons Ltd, 2005.

[7] Yanfang Ye · Dingding Wang · Tao Li · Dongyi Ye , "*An intelligent PE-malware detection system based on association mining*", J Comput Virol (2008) 4:323–334, Springer-Verlag France 2008.

[8] Adel Sabry Issa, "*A Comparative Study among Several Modified Intrusion Detection System Techniques*", Master Thesis, University of Duhok, 2009.

[9] Luai Al Shalabi, Zyad Shaaban and Basel Kasasbeh, "*Data Mining: A Preprocessing Engine*", Journal of Computer Science 2 (9): 735-739, 2006, ISSN 1549-3636, Science Publications, 2006.

[10] Paul K. Harmer, Paul D. Williams, Gregg H. Gunsch, and Gary B. Lamont, "*An Artificial Immune System Architecture for Computer Security Applications*", IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 6, NO. 3, JUNE 2002.

[11] Dipankar Dasgupta and Luis Fernando Niño, *"Immunological Computation Theory and Applications"*, book, 2009.

[12] Zhou Ji and Dipankar Dasgupta, *"Revisiting Negative Selection Algorithms"*, Massachusetts Institute of Technology, 2007.

[13] Thomas Stibor, *"On the Appropriateness of Negative Selection for Anomaly Detection and Network Intrusion Detection"*, PhD thesis 2006.

[14] Rune Schmidt Jensen, *"Immune System for Virus Detection and Elimination"*, IMM-thesis-2002.

[15] Fernando Esponda, Stephanie Forrest and Paul Helman, *"A Formal Framework for Positive and Negative Detection Schames"*, IEEE 2002.

[16] A. H. Momeni Azandaryani M. R. Meybodi, *"A Learning Automata Based Artificial Immune System for Data Classification"*, Proceedings of the 14th International CSI Computer Conference, IEEE 2009.

[17] Chao Chen, Andy Liaw, and Leo Breiman, *"Using Random Forest to Learn Imbalanced Data"*, Department of Statistics,UC Berkeley, 2004.

[18] Yuchun Tang, Sven Krasser, Paul Judge, and Yan-Qing Zhang, *"Fast and Effective Spam Sender Detection with Granular SVM on Highly Imbalanced Mail Server Behavior Data"*, (Invited Paper), Secure Computing Corporation, North Point Parkway, 2006.

[19] Jamie Twycross , Uwe Aickelin and Amanda Whitbrook, *"Detecting Anomalous Process Behaviour using Second Generation Artificial Immune Systems"*, University of Nottingham, UK, 2010.

[20] H. Güneş Kayacık, A. Nur Zincir-Heywood, and Malcolm I. Heywood, *"Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets"*, 6050 University Avenue, Halifax, Nova Scotia. B3H 1W5, 2006.

[21] Feng Gu, Julie Greensmith and Uwe Aickelin, *"Further Exploration of the Dendritic Cell Algorithm: Antigen Multiplier and Time Windows"*, University of Nottingham, UK, 2007.

[22] Prachya Pongaksorn, Thanawin Rakthanmanon, and Kitsana Waiyamai, "*DCR: Discretization using Class Information to Reduce Number of Intervals*", Data Analysis and Knowledge Discovery Laboratory (DAKDL), P. Lenca and S. Lallich (Eds.): QIMIE/PAKDD 2009.

[23] Chen Guangzhu, Li Zhishu, Yuan Daohua, Nimazhaxi and Zhai yusheng. "*An Immune Algorithm based on the Complement Activation Pathway"*, IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.1A, January 2006.

[24] J. McHugh, "Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory," *ACM Transactions on Information and System Security*, vol. 3, no. 4, pp. 262–294, 2000.

[25] *The NSL-KDD Data Set*, http://nsl.cs.unb.ca/NSL-KDD.

[26] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "*A Detailed Analysis of the KDD CUP 99 Data Set*," Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.

UML Model of Deeper Meaning Natural Language Translation System using Conceptual Dependency Based Internal Representation

Sandhia Valsala
College of Computer Studies
AMA International University
Salmabad,Kingdom of Bahrain
sandhia_v@yahoo.com

Dr Minerva Bunagan
College of Computer Studies
AMA International University
Salmabad,Kingdom of Bahrain
yarahph@yahoo.com

Roger Reyes
College of Computer Studies
AMA International University
Salmabad,Kingdom of Bahrain
roger.reyes@yahoo.com

***Abstract*—** Translation from one language to another language involves many mechanical rules or statistical inferences. Statistical inference based translations lack any depth or logical basis for the translation. For a deeper meaning translation to be performed using only the mechanical rules are not sufficient. There is a need to extract suggestions from common world knowledge and cultural knowledge. These suggestions can be used to fine tune or may be even reject the possible candidate sentences. This research presents a software design for a translation system that will examine sentences based on the syntax rules of the natural language. It will then construct an internal representation to store this knowledge. It can then annotate and fine tune the translation process by using the previously stored world knowledge.

Keywords

Natural language, Translation, Conceptual Dependency,Unified Modeling Language (UML)

I. Introduction

Living in an electronic age has increased international interaction among individuals and communities. Rapid and accurate translation from one natural language to another is the required for communication directly with individuals natives of a foreign language.

Automated translation desired by anyone wishing to study international subjects. There are a large number of naturally spoken languages. Some automated software systems are available that allow translation from one natural language to another. By using these systems one can translate a sentence from one natural language to another without any human translator. But these systems often fail to convey the deeper meaning of original text to the translated language.

The objective of this paper is to present a design an automated natural language translation system from English to Urdu or Arabic. This system will use a system-internal representation for storing the deeper meaning of input sentences. This paper will also identify natural language grammar rules that can be used to construct this system.

II. Definition of Terms

a.Natural Language

Natural language is any language used by people to communicate with other people. In this paper the two natural languages selected for translation are English and Urdu. The methods described here are generally extendable for most natural languages.

b.Grammar of a Natural Language

Grammar of a language is a set of production rules (Aho et al., 2006) using meta-symbols or non-terminals and tokens (class of words of the language). These rules can be used to determine if a sentence is valid or invalid. Extended Backus-Naur Form (EBNF) is used to theoretically describe such grammars (Rizvi, 2009) (Wang, 2009).

c. Conceptual Dependency

The theory of Conceptual Dependency (CD) was developed by Shank and his fellow researches for representing the higher level interpretation of natural language sentences and constructs (Shank and Tesler, 1969). It is a slot-and-filler data structure can be modeled in an object oriented programming language (Luger and Stubblefield, 1996). CD structures have been used as a means of internal representation of meaning of sentences in several language understanding systems (Schank and Riesbeck, 1981).

III. Review of Relevant Literature

Automated translation systems from companies like Google and Microsoft use probability and statistics to predict translation based upon previous training (Anthes, 2010). Usually they train on huge sample data sets of two or more natural language document sets. In a situation where there is a sentence using less commonly used words so that no translation exists previously for those group of words such a translation system may not give accurate results.

Conceptual Dependency (CD) theory has been developed to extract underlying knowledge from natural language input (Shank and Tesler, 1969). The extracted knowledge is stored and processed in the system using strong slot-and-filler type data abstractions. The significance of CD to this research is that it describes a natural language independent semantic network that can be used to disambiguate the meaning by comparing it with internally stored common world knowledge.

Conceptual dependency theory is based on a limited number of primitive act concepts (Schank and Riesbeck, 1981). These primitive act concepts represent the essence of the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

meaning of an input sentence and are independent of syntax related peculiarities of any one natural language. The important primitive acts are summarized in Table 1.

Table 1 - Schank's Primitive Act Concepts.

Primitive Act	Description	Example
ATRANS	Transfer of an abstract relationship such as possession, ownership or control. ATARNS requires an actor, object and recipient.	give, take, buy
PTRANS	Transfer of the physical location of an object. PTRANS requires an actor, object and direction.	go, fly
PROPEL	Application of a physical force to an object. Direction, object and actor are required.	push, pull
MTRANS	Transfer of mental information between or within an animal.	tell, remember, forget
MBUILD	Construction of new information from old information.	describe, answer, imagine
ATTEND	Focus a sense on a stimulus.	listen, watch
SPEAK	Utter a sound.	Say
GRASP	To hold an object.	Clutch
MOVE	Movement of a body part by owner.	kick, shake
INGEST	Ingest an object. It requires an actor and object.	Eat
EXPEL	To expel something from body.	

Valid combinations of the primitive acts are governed by 4 governing categories and 2 assisting categories (Schank and Tesler, 1969). These conceptual categories are like meta-rules about the primitive acts and they dictate how the primitive acts can be connected to form networks. In Schank and Tesler's work there is implicit English dependent interpretation of Producer Attribute (PA) and Action Attribute (AA). But in this research the interpretation of PA and AA is natural language independent. The conceptual categories are summarized in Table 2.

Table 2 - Schank's Conceptual Categories.

Governing Categories	
Name	**Description**
PP	Picture Producer. Represents physical objects
ACT	Action. Physical actions.
LOC	Location. A location of a conceptualization.
T	Time. Time of conceptualization.
Assisting Categories	
Name	**Description**
PA	Producer Attribute. Attribute of a PP.
AA	Action Attribute. Attribute of an ACT.

Traditionally EBNF grammar rules are used to express a language grammar (Aho et al., 2004). Most natural languages in general and English in particular has been a particular focus of research in many countries (Wang, 2009). A study of the Urdu language grammar for computer based software processing has been done previously (Rizvi, 2007). Urdu language shares many traits with Arabic and other South-Asian languages. Traits like common script and some common vocabulary are the most well known of these.

IV. Implementation

Materials and Methods

For the purpose of design of the software this research utilizes English as the first or source natural language and Urdu as the second or target natural language. This choice is based primarily upon the familiarity of the researchers with the languages. Another reason is that EBNF grammar is available for these languages (Wang, 2009) (Rizvi, 2007). However, the design presented here can be equally appropriate for most of the natural languages. The design primarily uses UML diagrams notation and can be drawn in Microsoft Visual Studio 2010 (Loton, 2010) or Oracle JDeveloper software (Miles and Hamilton, 2006).

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The design is broken into two main use-case scenarios. The first use-case is for first natural language user (English). The system components identified in this use case include a tokenizer, parser, CD annotator and CD world-knowledge integrator. In this use-case the working system will take an input sentence and then construct an internal representation of that sentence. The user will be returned a Reference ID (REFID) number which is a mechanism to identify the internal representation (concept) inside the systems memory. The second use-case is for the target language user (Urdu). The user identifies an internal concept through a REFID. The system will then generate the corresponding Urdu sentence. The system components identified in this use-case include CD world-knowledge integrator, tokenizer and sentence formulator. Two sequence diagrams corresponding to the two use cases are shown in Figure 1 and Figure 2.

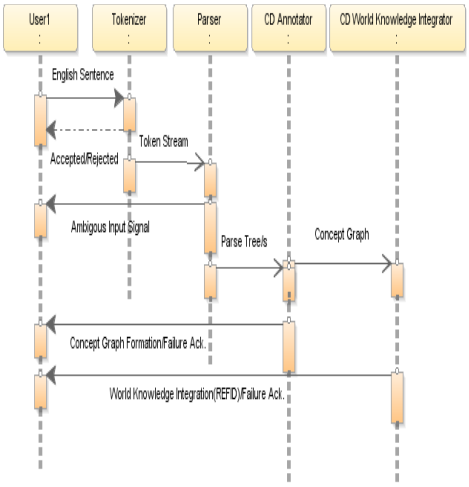


Figure 1 - Sequence diagram for User Input Language Processing use case.

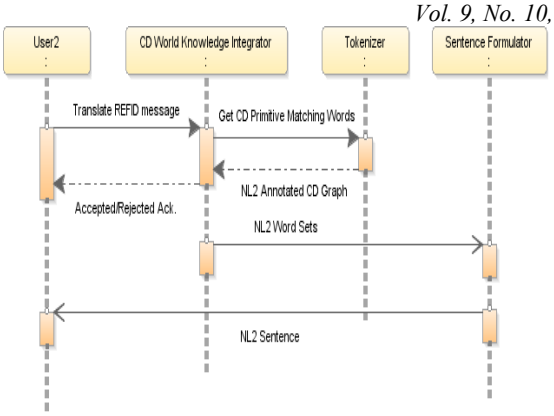


Figure 2 - Sequence diagram for target natural language conversion use case.

A discussion of the functions of the major components identified in these figures is given below.

Tokenizer

Tokenizer component will have two functions. The first function will take a source natural language sentence as input and it will create a stream of tokens from it if the words are found in the dictionary of the language. Tokens can be an extension of the parts of speech of the natural language (English) or taken from the terminal symbols in the EBNF grammar. These tokens will be used in specifying the EBNF grammar rules. This function will also generate an Accepted or Rejected signal for the User. If the token stream is valid it will be passed to the Parser component. This function is shown in Figure 1.

The second function of the tokenizer component is in target natural language conversion use case. This function will take input of a CD primitives graph and return all corresponding words found in the dictionary of the target natural language. Tokenizer component can be implemented in an object oriented programming language. This function is shown in Figure 2.

Parser

Parser component will take as input a token stream consisting of tokens from the source natural language parts of speech or grammar terminal symbols. The parser will match the token stream against all syntax rules of the source natural language. If the sentence is valid and unambiguous one parse tree will be generated as output. If the sentence is not valid an error message will be given as output. If the sentence is ambiguous then all parse trees will be returned to the calling component for a possible selection. The selected parse tree will be given as input to the CD Annotator component for further processing. This component is shown in context in Figure 1. For most natural languages the parser component can be

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

prototyped or implemented in Prolog programming language and it may be generated from a LR parser generator tool like YACC or Bison.

CD Annotator

CD annotator component will take as input the parse tree generated by the parser component and create and annotate a CD graph data structure. The CD graph structure will be based upon the CD primitives as listed in Table 1 and Table 2. The CD graph data structure can be implemented in an object oriented programming language. This component is shown in Figure 1.

CD World Knowledge Integrator

This component will have two main functions. First of all it will add the new sentence Concept Graph into a bigger common world knowledge graph. The common world knowledge will consist of facts like "Gravity pulls matter down", "Air is lighter than water", etc. This knowledge will be relevant to the closed world assumption of a Faculty Room in the University. Internally this knowledge will be represented in CD form itself. Upon receiving new input this component will create links with common world knowledge already stored in the system. After integration of the new Concept Graph a Reference Identification number (REFID) will be returned to the user for later retrieval of the newly stored concept. This function is shown in Figure 1.

Second function of this component will be to receive as input a REFID number and to locate its corresponding integrated concept graph. By scanning the integrated concept graph it will generate a list of primitive CD in use in the REFID referenced integrated concept graph. This list will be passed to the tokenizer component which will return target natural language word sets matching the list of primitive CD. These word sets will be used by this component to annotate the integrated concept graph with target natural language words. The target natural language annotated CD graph will be given as input to sentence formulator component for sentence generation. This function is shown in Figure 2.

Sentence Formulator

Sentence Formulator component will take as input the target natural language annotated CD graph and it will apply the syntax rules of the target language to produce valid sentences of the target language. This component is shown in Figure 2.

Design of the Parser

This research presents a simple Prolog Programming Language English parser (Appendix), that is based on the English grammar rules described in (Wang, 2009) and as taught in university English courses.

Conceptual Dependency Graph

In this research CD based object oriented (OO) architecture is proposed for the internal representation of meaning of the natural language. Each primitive concept has to be implemented as a class in an OO programming language. Most of these classes will have predefined attributes and some implementation specific attributes will be added to them. The work done by (Schank and Tesler, 1969) provides general rules concerning the structure and meaning of such a network.

Language Dictionaries

For the source natural language and the target natural language a Dictionary will have to be created. It can be implemented as a file or a database. The dictionary will contain words from the closed world scenario (Faculty Room). For each word part-of-speech attribute (or the corresponding EBNF non-terminal symbol name) will have to be identified. For some words there will also be mappings to primitive concepts (Table 1).

English Grammar in Prolog Programming Language

The following computer program is a source-code listing in Prolog Programming Language. It describes a simple English sentence parser. It can validate or invalidate a sentence made of words in the vocabulary. For testing purposes, this parser can be used to generate sentences of a given word length according to the words in vocabulary and Prolog unification order. It has been tested on SWI Prolog Programming Environment. (http://www.swi-prolog.org)

```
/* **** English Sentence Grammar in Prolog  */
/*      Assumes a closed world assumption       */
/*      Faculty room in a university            */
/* ****************************** */

/* In absence of a Tokenizer, hard coding of words
(vocabulary) and Tokens */
p_noun('pname1').

imp_noun('student').
imp_noun('book').

pro_noun_subject('i').
pro_noun_subject('he').
pro_noun_subject('she').
pro_noun_subject('we').
pro_noun_subject('they').
pro_noun_subject('it').

pro_noun_object('me').
pro_noun_object('him').
pro_noun_object('her').
pro_noun_object('us').
pro_noun_object('them').
pro_noun_object('it').
```


http://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
pro_noun_possesive('his').
pro_noun_possesive('her').
pro_noun_possesive('their').
pro_noun_possesive('our').
pro_noun_possesive('your').
pro_noun_possesive('whose').

pro_noun_nominative_possesive('mine').
pro_noun_nominative_possesive('yours').
pro_noun_nominative_possesive('ours').
pro_noun_nominative_possesive('theirs').

pro_noun_indefinite('few').
pro_noun_indefinite('more').
pro_noun_indefinite('each').
pro_noun_indefinite('every').
pro_noun_indefinite('either').
pro_noun_indefinite('all').
pro_noun_indefinite('both').
pro_noun_indefinite('some').
pro_noun_indefinite('any').

pro_noun_demonstrative('this').
pro_noun_demonstrative('that').
pro_noun_demonstrative('these').
pro_noun_demonstrative('those').
pro_noun_demonstrative('such').

/* For ease in testing reducing the number of unifications,
limited items defined */
person('pname1').
person('student').

thing('book').

verb('sings').
verb('teaches').
verb('writes').

adjective('thick').
adjective('brilliant').

preposition('in').
preposition('on').
preposition('between').
preposition('after').

article('a').
article('an').
article('the').

/* Actual Rules */
noun(X) :- p_noun(X).
noun(X) :- imp_noun(X).

sub_noun(X) :- pro_noun_subject(X).
sub_noun(X) :- noun(X), person(X).

obj_noun(X) :- pro_noun_object(X).
obj_noun(X) :- pro_noun_nominative_possesive(X).
obj_noun(X) :- noun(X).

subject(X) :- sub_noun(X).

object(X) :- obj_noun(X).

indirect_object(X) :- pro_noun_object(X).
indirect_object(X) :- noun(X), person(X).

determiner(X) :- article(X).
determiner(X) :- pro_noun_possesive(X).
determiner(X) :- pro_noun_indefinite(X).
determiner(X) :- pro_noun_demonstrative(X).

noun_phrase(X) :- noun(X).
noun_phrase([X|Y]) :- adjective(X), listsplit(Y, H, T), T=[],
noun(H).

preposition_phrase([X|Y]) :- preposition(X), listsplit(Y, H1,
T1), determiner(H1), noun_phrase(T1).

object_complement(X) :- noun_phrase(X).
object_complement(X) :- preposition_phrase(X).
%% object_complement(X) :- adjective_phrase(X).

/* Breaking the head off a list */
listsplit([Head|Tail], Head, Tail).

/* Determining length of list */
listlength([], 0).
listlength([_|Y], N) :- listlength(Y, N1), N is N1 + 1.

/* Pattern1: Subject-Verb */
sentence([X|Y]) :- subject(X), listsplit(Y, Head, Tail), Tail=[],
verb(Head).

/* Pattern2: Subject-Verb-Object */
sentence([X|Y]) :- subject(X), listsplit(Y, H, T), verb(H),
listsplit(T, H2, T2),
                object(H2), T2=[].
sentence([X|Y]) :- subject(X), listsplit(Y, H, T), verb(H),
listsplit(T, H2, T2),
                pro_noun_possesive(H2), listsplit(T2, H3, T3),
                        object(H3), T3=[].
/* Pattern3: Subject-Verb-Indirect Object-Object */
sentence([X|Y]) :- subject(X), listsplit(Y, H, T), verb(H),
                        listsplit(T, H2, T2),
                indirect_object(H2), listsplit(T2, H3, T3),
                        object(H3), T3=[].
sentence([X|Y]) :- subject(X), listsplit(Y, H, T), verb(H),
                        listsplit(T, H2, T2),
                indirect_object(H2), listsplit(T2, H3, T3),
                pro_noun_possesive(H3), listsplit(T3, H4, T4),
```


http://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
                    object(H4), T4=[].
/* Pattern4: Subject-Verb-Object-Object Complement */
sentence([X|Y]) :- subject(X), listsplit(Y, H, T), verb(H),
listsplit(T, H2, T2),
                    object(H2), object_complement(T2).
```

V.Conclusion and Recommendations

A system level modular design of a software system for translation between a source natural language to a target natural language was presented. A functional behaviour of each of the major software components was also discussed.

For extending this system to other languages the following 3 additions will need to be made. First of all an EBNF grammar should be made available for new language to be integrated. Second a system dictionary should be created for the new language as mentioned above. And third, the tokenizer, parser and sentence formulator components need to be enhanced to handle the new language. These components form the front-end (user facing part) of the system. The back end remains unchanged.

For extending the scope of the system translation from the closed-world-scenario of a faculty room to more general translator, universal common knowledge base can be integrated into this system design. One such universal common knowledge base is the CYC project as described in (Lenat et al., 1990).

VI. References

1.Aho, Alfred V., Lam, Monica S., Sethi, Ravi, and Ullman, Jeffery D. (2006) *Compilers: Principles, Techniques, and Tools*. Addison Wesley Publishing Company, 1000pp.

2.Anthes, Gary (2010) Automated Translation of Indian Languages. *Communications of the ACM,* Vol 53, No. 1: 24-26

3.Lenat, Douglas B., Guha, R. V., Pittman, Karen, Pratt, Dexter, and Shepherd, Mary (1990), Cyc: toward programs with common sense, *Communications of the ACM,* Volume 33, Issue 8

4.Loton, Tony (2010), *UML Software Design with Visual Studio 2010: What you need to know, and no more!* CreateSpace Press, 136pp

5.Luger, G. and Stubblefield, W. (1997) *Artificial Intelligence: Structures and Strategies for Complex Problem Solving.* Addison Wesley Publishing Company, 868pp.

6.Miles, Russ and Hamilton, Kim (2006), *Learning UML 2.0*, O'Reilly Media, 288pp

7.Rizvi, Syed M. J. (2007) *Development of Algorithms and Computational Grammar for Urdu.* Doctoral Thesis, Pakistan Institute of Engineering and Applied Sciences Nilore Islamabad, 242pp

8.Schank, Roger C. and Tesler, Larry (1969) A Conceptual Dependency Parser for Natural Language. *Proceedings of the 1969 conference on Computational linguistics Association for Computational Linguistics Stroudsburg, PA, USA*. doi:10.3115/990403.990405

9.Schank, Roger C. and Riesbeck, Christopher K. eds. (1981) *Inside Computer Understanding: Five Programs plus Miniatures.* Psychology Press, 400pp, http://www.questia.com Web.

10.Wang, Yingxu (2009) A Formal Syntax of Natural Languages and the Deductive Grammar. *Journal Fundamenta Informaticae - Cognitive Informatics, Cognitive Computing, and Their Denotational Mathematical Foundations (II)*, Vol 90, Issue 4

AUTHOR'S PROFILE:

Ms.Sandhia Valsala, is presently associated with AMA International University, Bahrain as Asst Professor in the Computer Science Department. She holds a Master's degree in Computer Applications from Bharatiyar University, Coimbatore and is currently pursuing her Phd from Karpagam University Coimbatore.

Dr Minerva Bunagan is presently associated with AMA International University,Bahrain as the Dean of College of Computer Studies .She holds a P.hD in Education from Cagayan State University, Tuguegarao City, Philippines. She also holds a Master of Science in Information Technology from Saint Paul University Philippines8.

Roger Reyes is presently associated with AMA International University, Bahrain as Asst Professor in the Computer Science Department. He holds a masters degree from AMA computer university
Quezon city Philippines

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Monitoring Software Product Process Metrics

Zahra Gholami
Department of Software Engineering
North Tehran Branch - Islamic Azad
Tehran, Iran
Gholamizahra64@yahoo.com

Nasser Modiri
Department of Software Engineering
Zanjan Branch - Islamic Azad
Zanjan, Iran
NasserModiri@yahoo.com

Sam Jabbedari
Department of Software Engineering
North Tehran Branch - Islamic Azad
Tehran, Iran
S_jabbehdari@iau-tnb.ac.ir

Abstract— Software quality is an important criterion in producing softwares which increases productivity and results in powerful and invincible softwares. We can say that quality assurance is the main principle and plan in software production. One of the most important challenges in Software Engineering is lack of software metrics for monitoring and measurement of software life cycle phases which causes low quality and usefulness of software products. Considering the importance of software metrics, utilization of international standard software life cycle process model (ISO/IEC 12207) and measurement process of Plan/Do/Check/Act in order to monitor software production cycle is presented in this paper.

Keywords-Software Metrics, Measurement, Software Product Process, ISO/IEC 12207

I. INTRODUCTION

Nowadays development and quality improvement of software production process and increasing performance and throughput of people involved is an important matter for every corporation which deals with information technology and software industry. Requests for efficient software have increased since computers became more powerful and because of vital role of technology in promotion of business, software problem are effective on most companies and governments. These days many companies realized that most of software problems are technical and software engineering is different from other engineering fields because software products are intellectual but the other engineering products are physical. There is measurement in centre of every engineering which is a method based on known standards and agreements. Software metrics include wild range of measurements for computer softwares, also measurement could be used throughout the software project in order to help estimation, quality control, throughput evaluation, project control. The main aim of this essay is to review and propose parameters as software metrics which are applied in standard ISO/IEC 12207 in order to remove weakness points of this standard and also helping us in quality measure of mentioned standard and to provide the possibility of quality effective factors investigation in software product process [9].

II. SOFTWARE PRODUCT PROCESS

Software product process is a structure and also a framework for introducing organization in order to design and generate a new software product consist of key solutions, issues and problems of a software product from early stages of marketing to mass production and finally release that[6].

Figure1. Software Product Process

III. SOFTWARE METRICS

Software metrics are parameters for measuring softwares which measurement won't have any meaning without them. It doesn't mean that software metrics can solve every problem but they can conduct managers to improve processes, throughput and quality of softwares [4]. Metrics are continuous and executable activities on whole project and are collected in long period of time; they show the rate of progress in periodic performances. Metrics have ring-incremental mechanism because the most valuable information is obtained when we have a sequence of data. Then the data obtained from metrics as feedback should have been given to the manager in order to find existing mistakes, provide solution for them and prevent further rising of faults. This makes defects detection be done before presentation to the customer.

A. *Metrics Types*

Other metrics can be defined with considering different viewpoints such as:

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

1) Subjective Metrics

These metrics can't be evaluated and are express with a set of qualitative attributes. The main objective of these metrics is to identify and evaluate of metrics which are less ponderable quantitatively.

2) Objective Metrics

Metrics that can be evaluated and are measurable such as number of human resources, number of resources, size of memory, number of documentation and number of modules.

3) Global Metrics

These metrics are used by software managers and are comprehensive metrics which we can evaluate project status with using of them, such as the budget, project time schedule, cost of implementation.

4) Phase Metrics

These kinds of metrics are specific to each phase and they measure the rate of progress or regression in specific phase. For example number of people in each phase, specific documentation of phase, improvement percent, and delay percent.

5) Calculated Metrics

These metrics can be calculated. For example cost, error, complexity, rate of execution, execution time.

6) Product Metrics

These are metrics that analyze final product for example the time needed for presentation of product, rate of execution, maintenance costs, and product user friendliness.

7) Resource Metrics

Metrics which describe feature of available resources. For example number of programmers, analysts, designers and required systems.

8) Risk Metrics

Metrics that are used for identification, giving priority to the probable risks of projects and reducing the probability of them.

9) Management Metrics

Metrics that are used for progress and development of project management [1, 2, 3, 8].

IV. PLAN/DO/CHECK/ACT

Plan/Do/Check/Act Cycle was established by Japanese in 1951 based on doming cycle. This cycle consist of four following stages:

Plan: determining of objectives and required process for presentation of results according to customer's requests and or organization policies.

Do: implementation

Check: monitoring and measurement of process and product according to policies, objectives and requirements or request related to product and reporting of results.

Act: doing activities in order to improve process performance.

This cycle is based on scientific methods and feedback plays a basic role in that so the main principle of this scientific method is iteration. When a hypothesis is being denied the next execution of cycle can expand knowledge and these iteration makes become closer to the aim. A Process is partitioned into PDCA Activities show in Figure2 [5].

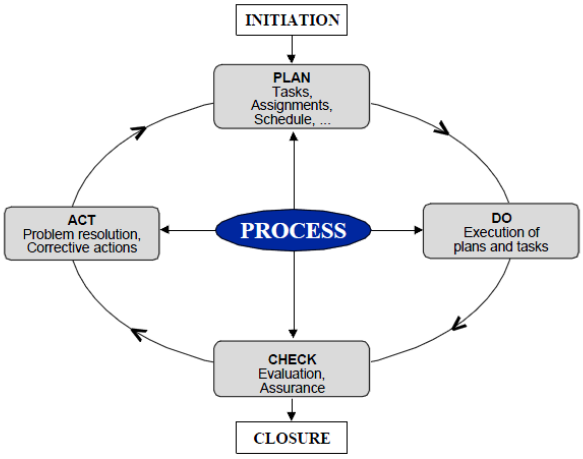


Figure 2. Partitioning a Process into PCDA Activities [7]

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. Proposed of Software Metrics Cycle according to Plan/Do/Check/Act

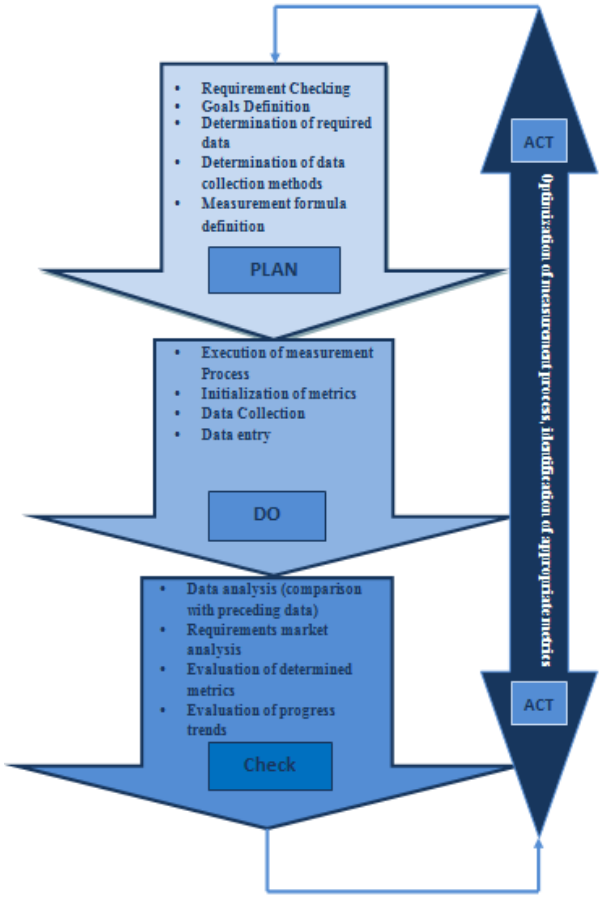


Figure3. Software metrics cycle according to plan/Do/Check/Act

V. Proposed Pattern

Considering the point that plan/Do/Check/Act is a simple and effective process for measurement of software metrics following of that is a high assurance for success in control and monitoring of software production cycle metrics and with considering the weaknesses of standard ISO/IEC 12207 we can apply our desired metrics during this cycle to the different phases of the mentioned standard so that defects would be eliminated to some extent (Figure 4).

A. Features of Pattern

In this pattern we apply different metrics considering the importance via plan/Do/check/Act cycle and features which we can express for this pattern consist of followings:

1) Further Reliability

With using resource, risk and management metrics which are the most important metrics at the start of project and utilization of Plan/Do/Check/Act cycle for each metrics we can provide further monitoring and control on production processes and so further reliability for establishing a project will be realized.

2) Cost Reduction

With using metrics which are applied to the standard ISO/IEC 12207 we can prevent next duplication because of observation at the start of project.

3) Risk Reduction

We can also minimize the risk with using risk and management metrics.

VI. Conclusion

Result of this essay is proposal of a pattern that is based on standard ISO/IEC 12207 and uses proposed metrics for monitoring of processes. One of the methods for controlling and monitoring of software production process is software metrics that can be applied to every phase so that transition to the next phase would be more assured. It should be noted that this reliability isn't completely definite but it can prevent increasing cost because of negligence to some parameters so metrics are necessary and essential.

References

[1] N. Fenton, and S. Pfleeger, "Software Metrics - A Rigorous and Practical Approach," Brooks Cole Publishing Company, ISBN: 0534954291, 1998.

[2] Somerville, "Software Engineering," Pearson Education, ISBN: o321210263, 1999.

[3] R. Pressman, "Software Engineering: A Practitioner's Approach 5th Edition," McGraw-Hill Science/Engineering/Math, ISBN: 0072853182, 2001.

[4] P. Goodman "Software Metrics: Best Practices for Successful IT Management," Rothstein Associates Inc., 2004.

[5] G. Gorenflo, J. W. Moran, "The ABCs of PDCA for Public Health Foundation's website, 2010.

[6]Day, G. "The Product Life cycle: Analysis and applications issues," *Journal of Marketing*, Vol 45, autumn 1981.

[7] Raghu Singh, "An Introduction to International Standard ISO/IEC 12207 Software Life Cycle," FAA, Washington DC, April 26, 1999.

[8] A Guide to the Project Management Body of Knowledge (PMBOK), an American National Standard, ANSI/PMI 99-001-2008, 2008.

[9] ISO/IEC 12207:2008, International Standard, System and software engineering-Software life cycle processes.

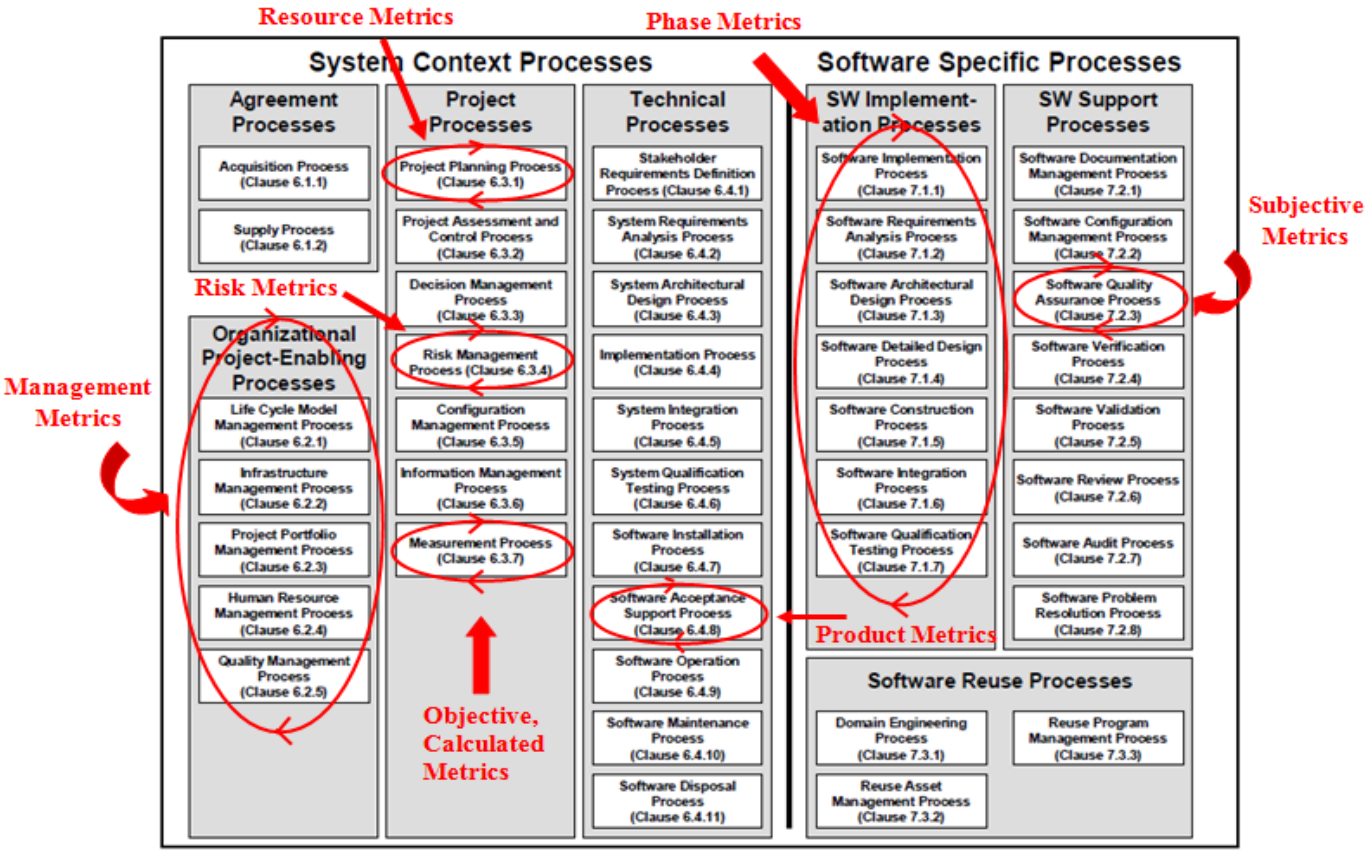


Figure 4. Proposed Pattern [9]

AUTHORS PROFILE

Zahra Gholami received her B.Sc Degree in Computer Engineering from Islamic Azad University (Lahijan branch), Lahijan, Iran in the year 2008. Currently she is pursuing M.Sc in Computer Engineering from Islamic Azad University (Tehran North Branch), Tehran, Iran under the guidance of Dr Modiri. She is presently working on Metrics of Software Development Process.

Nasser Modiri received his M.S. degree from the University Of Southampton, U.K., and Ph. D. degree from the University of Sussex, U.K. in 1986 and 1989, respectively. In 1988 he joined The Networking Centre of Hemel Hempstead, and in 1989 he worked as a Principal Engineer at System Telephone Company (STC) Telecommunications Systems, U.K.

Currently, Dr. Modiri is the president of Ayandehgan Rayaneh Co. developing web-based software and designer and implementer of information technologies services for Intranet networks while teaching actively MSc courses in network designing, software engineering and undertaking many MSc projects. He is currently developing applications for Virtual Universities, Virtual Parliaments, Virtual Organizations, ERP, GPS+GSM, GPRS, RFID, ISO/IEC 27000, ISO/IEC 15408 technologies.

Sam Jabbedari received his both B.Sc. and M.S. degrees in Electrical Engineering Telecommunication from K.N.T (Khajeh Nasir Toosi) University of Technology, and IAU, South Tehran branch in Tehran, Iran, in 1988, through 1991 respectively. He was honored Ph.D. Degree in Computer Engineering from IAU, Science and Research Branch, Tehran, Iran in 2005.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Designing a Comprehensive Model for Evaluating SOA-based Services Maintainability

Maryam Zarrin
Computer Engineering Department,
Science & Research Branch of Islamic Azad University,
Tehran, Iran
zarinzeus@gmail.com

Mir Ali Seyyedi
Computer Engineering Department,
Islamic Azad University, Tehran-south branch
Tehran, Iran
Maseyyedi2002@yahoo.com

Mehran Mohsenzaeh
Computer Engineering Department,
Science & Research Branch of Islamic Azad University,
Tehran, Iran
M_mohsenzadeh77@yahoo.com

***Abstract*- The aim of this paper is to propose a comprehensive and practical model to evaluate the maintainability of software services in service-oriented architecture in the entire service lifecycle and based on the fuzzy system. This model provides the possibility of making decisions concerning the maintainability of SOA-based services for service managers and owners in various service operation and design phases. The proposed maintainability evaluation model consists of five sections: input, analysis, measurement, decision making and output. According to the studies conducted on the leading article, the service structural properties in design phase as well as service management mechanism structures at the operation phase have been identified as effective factors in evaluating the maintainability of services. So the proposed model investigates both discussed factors and is generally categorized in two sections: design and operation. To assess maintainability in both sections, the fuzzy technique is used.**

Keywords- maintainability; service-oriented; evaluation model; fuzzy system

I. INTRODUCTION

In recent years, the use of service-oriented architecture as one of the significant solutions for managing complexities and interactions between IT-based services as well as managing fast business shifts in a volatile business environment has increased. Maintainability is one of the major service quality attributes which has an important role in user satisfaction and cost reduction of maintenance and support. Research has shown that more than 60% of the overall resources devoted to software or services development belongs to the maintenance phase [21]. So, designing services that face a difficulty at the maintenance phase will greatly increase the possibility of cost or time failure of service development [21].

According to the definition provided by IEEE, maintainability is defined as a capability of the software against possible adjustments like correcting errors, improving efficiency or other software quality attributes or adaptation of the software with the environment, functionality or requirement changes [14]. Also, to measure whether an IT service or component configuration after encountering failure in service maintainability area, how quickly and effectively could return to its normal activity is a description that has been presented by the third version of ITIL standards [15].

Presently, little research effort has been dedicated to considering maintainability evaluation of SOA-based services and more significantly, practical model for evaluating maintainability of service-oriented services regarding all maintainability influencing factors in entire service lifecycle do not exist. In other words, the focus of researches of the existing models has been more on maintainability evaluation and assessment in the software perspective.

Due to the service-oriented architecture characteristics and their differences with others, the factors and metrics used in these models have not been applicable in service-oriented approaches and they are not directly functional in the service orientation perspective. So in recent years, studies on maintainability evaluation have been conducted in order to establish and define appropriate metrics and models in service orientation context. Nonetheless, the study conducted in this area is at research and theory level which has been investigated in limited dimensions and also a comprehensive and practical method for evaluating the SOA-based service maintainability has not been presented. The only researches presented in this area include two evaluation models that have been presented by Mikhail Perepletchikov [3, 5]. Linear regression prediction models have been used in both models but in the first, coupling metrics have been presented [4] and in the second, the cohesion metrics [6] have been used as model predictors.

Other existing researches in this context are limited to proposing new metrics for evaluating the services design structural properties. So far by using these metrics, comprehensive and practical models for evaluating the services maintainability in service-oriented approach have not been introduced. In both [19, 20] researches metrics have been proposed to evaluate decoupling using connections

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

between components based on service orientation as well as in [10] dynamic metrics of coupling measurement with regards to run time relations between services. [11, 12] includes a set of metrics to measure the complexity of the service-oriented design systems. Also in [9] considering principles such as loose coupling and appropriate granularity in designing services, metrics for them have been proposed. In [8] reusability, composability, granularity, cohesion and coupling evaluation metrics by available information in service-oriented design have been proposed.

Obviously, the comprehensive evaluation of maintainability in service-oriented architecture will have a perception in the service lifecycle. In other words, designing and defining a comprehensive model for evaluating SOA-based services maintainability will be possible by considering maintainability influencing factors in the full service lifecycle. By having such model, senior managers and service owners will be able to make decisions on maintainability of SOA-based services not only at every stage of the service design and operation but also when services are operational.

This paper makes a contribution in proposing a comprehensive and practical model for evaluating the maintainability of SOA-based services covering all maintainability influencing factors in full service lifecycle. The proposed evaluation model includes five sections: input, analysis, measurement, decision making and output.

In designing the evaluation model, the concept of maintainability is based on the included definitions and concepts in ITIL and four sub-attributes of ISO/IEC 9126 standards namely analyzability, changeability, stability and testability. It has also been considered as a combination of maintainability due to service structural properties in the design phase and operational phase of the service. As a result, evaluating the maintainability of services is conducted in two sections: one belonging to service design and the other to service operation phase factors.

In design sections, structural characteristics such as coupling, cohesion, and granularity directly affect the maintainability sub-attributes and indirectly service maintainability in which their effects can be estimated and predicted. Furthermore in the operation section, ITIL service management processes include incident management, problem management, change management, configuration management, release management and availability management which can directly map the maintainability sub-attributes, have a direct impact on maintainability.

Further, initially the model design requirements will be defined then methods and techniques used to answer each one of them will be provided. And at the end, maintainability evaluation proposed model will be described by using the fuzzy system and its various components.

II. PROBLEM DEFINITION AND APPROACH

To design the maintainability evaluation model in service-oriented architecture, it is first necessary to identify the SOA fundamental characteristics in relation to the previous architectural styles and identify their affects on model structural design. The next thing is to define the concept of maintainability in the entire service lifecycle. According to identification of the two phases of service design and operation as the major phases having influence on services maintainability and its minimal effect of other phases on it, it would be sufficient to define maintainability in the two introduced phases.

In other words, in the service design phase, it is necessary to divide the maintainability concept into the quality sub-attribute based on the available standards. Then the appropriate and associated factors must be determined from them. In next level, identification and selection of appropriate and associated metrics to every one of these factors is considered another challenge in designing this model. Also in the operation phase, initially the concept of maintainability should be divided into appropriate sub-attribute. Then, based on international standards each one of them must be mapped out based on appropriate process factors and the final step, the maturity level of every one of these process factors should be evaluated through certain metrics. In other words, the maintainability evaluation model should be defined in both service design and operation phases.

After determining the independent variables of the two phases, identifying their affects and significance on the maintainability dependent variable is an important challenge which an appropriate solution should be adopted for it. In the design phase, maintainability as a dependent variable and cohesion, coupling and granularity factors as independent variables are considered. So, in the first step of this phase, it is necessary to determine and identify the communications, affects and significance of each independent variable versus the dependent variable. In the next step through appropriate evaluation metrics selection service maintainability must be evaluated. Also in the operation phase, similar to the previous one determine the impact and the significance of each one of the independent variables on dependent variables and linked metrics is a major challenge that an appropriate solution for it should be adopted. In this section, service maintainability as dependent variable and the supportive process based on service management standards as independent variables are considered. Here, metric selection and efficient methods to evaluate process maturity level is another important challenge in this study that different aspects of it must be answered.

Another issue in designing this model is the selection of a metric evaluation technique or method from among methods used in other similar research or studies. In selecting an evaluation method measures such as compatibility with new data, viewing the reasoning process, suitability to complex models and also emphasis on compatibility with service-oriented architecture characteristics namely reusability, business agility, interoperability, loosely coupling and compos ability is important. Rest sections contribute to offer solutions to each of discussed areas.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. *Maintainability evaluation factors*

In services design phase, documentation-related factors and structural properties of design are significant influencing factors in term of maintainability evaluation. Documentation-related factors impact on maintainability are minimal because the proper of documentation increase the ability to analyze the failure in system or analyzability sub-attribute but it doesn't affect on service changeability and stability sub-attribute [2]. But according to the research conducted in the past, the structural properties which reflect the internal properties of services have a direct affect on all aspects of maintainability [22, 23 and 24]. As a result, if the structural properties of the product are appropriate, maintenance activities will simply be carried out. Thus, Documentation-related factors will be completely eliminated from the selection ones.

General structural properties of services include coupling, cohesion, size and complexity And SOA-specific structural properties services including service granularity, parameter granularity and consumability [1]. Selective structural properties include coupling, cohesion and granularity of the service. Complexity has been eliminated arguing the complexity of the design phase can be viewed as the combination of coupling and cohesion and in fact complexity is in a way duplicating two discussed properties [25]. The reason for eliminating the size by using a similar argument is the coverage of this feature with service granularity. Also the parameter granularity and consumability have been eliminated by documenting the shortage of their suggested sources as maintainability influence factors and as a result their minimal effect is overlooked. Therefore in design phase, maintainability is considered as a dependent variable and granularity, coupling and cohesion factors as independents.

In the operational phase, based on ISO/IEC 9126 standard maintainability was divided into four sub-attribute of analyzability, changeability, stability and testability [27]. Furthermore, for selecting the appropriate factors related to the sub-attributes, various service management standards such as ITIL and COBIT were evaluated. According to the purpose of this model, international ITIL framework that consists of two main areas of support and delivery were selected. ITIL framework focuses more on operational and tactical levels of service support and also includes effective procedures and processes to support services.

Efficient services managements depend on four areas: processes, products, people and provider. In other words, for optimal service management in the ITIL standard these four areas need to be properly assessed and evaluated. Further, by mapping the ITIL standard processes in the support area with maintainability sub-attribute, related and appropriate process according to table 1 were identified. So in this phase dependent variable is service maintainability and independent variables are support process levels include incident management, problem management, change management, configuration management, release management and availability management.

TABLE I. OPERATIONAL INDEPENDENT VARIABLES

ISO/IEC 9126 sub-attribute	*Appropriate processes of ITIL*
analyzability	incident management, problem management
changeability	change management, configuration management
stability	availability management
testability	release management

It should be noted in evaluation model designing, the addition of maintainability sub-attribute of ISO/IEC 9126 standard has been omitted because the addition of the mentioned level would increase the complexity and error of this model. So the considered category is solely for a better and more precise selection of suitable and related processes.

B. *The selection of metrics for maintainability evaluation factors*

Another challenge for this research is the selection of suitable metrics for evaluating maintainability factors which belong to the two phases of service design and operation. In services design phase, studies and research in the software and service-oriented metrics were studied. Overall, two metric categories were identified: 1) service-oriented specific metrics 2) software specific metrics. In the service-oriented architecture, the metrics related to structural properties are completely different from the software metrics [26]. Therefore, these types of metrics were completely eliminated.

Further, by using GQM technique and by accentuating on service-oriented architecture characteristics in terms of GQM components include Purpose, Aspects, Subject and Viewpoint, the appropriate questions were defined and based on them, the appropriate metrics of evaluating coupling [10], cohesion [1] and granularity [1] factors were chosen. Table 2 exhibits the selection metrics for the design phase.

TABLE II. EVALUATION METRICS FOR MAINTAINABILITY FACTORS

Structure property	*Complete name*	*metric*
coupling	Degree of Coupling within a given set of services metric (DCSS)	$DCSS = \frac{Max - \sum_{ueV}\sum_{veV} d(u,v)}{Max - Min}$ Max = K*V*(V-1) Max only appears when all of nodes in graph do not connect together Min = V*(V-1) Min only appears when all of nodes in graph connect to others.
cohesion	Inverse of Average Number of Used Message (IAUM)	$IAUM = \frac{SSNS}{TMU}$ SSNS: System Size in Number of Services TMU: Total Number of Message Used

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Structure property	*Complete name*	*metric*
Granularity	Squared Avg. Number of Operations to Squared Avg. Number of Messages	$AOMR = \frac{(\frac{NAO + NSO}{SSNS})2}{(\frac{TMU}{SSNS})2}$ NAO: Number of Asynchronous Operations NSO: Number of Synchronous Operations SSNS: System Size in Number of Services TMU: Total Number of Message Used

In the operational phase, due to the inefficiency of the GQM method in selecting the appropriate metrics, such as a lack of comprehensive questions in the method, self assessment techniques of OGC (the Office of Government Commerce) has been used as evaluation metrics for the operation phase [28]. This method includes a questionnaire that consists of all four dimensions of services management and evaluates them in nine levels through a variety of questions. Maturity level of selection process factors include prerequisites, management intent, process capability, internal integration, products, quality control, information management, integration and external interface with the customer.

C. Evaluation method

In a vast view, the proposed model with the modulation of design phase as well as service operations creates a maintainability evaluation structure. In this model, as for the offered evaluation structure, to provide a clear and unified response to, an evaluation technique is needed. Similar research and studies on prediction methods and quality characteristics were investigated [16, 17, 18, 7 and 13]. Generally two methods for predicting maintainability were identified: 1) Algorithmic technique model and 2) Hierarchical dimensional assessment model. To achieve the relationship function between independent and dependent variables, in the first batch from existing data set and in the second batch from expert opinions, probabilistic models and soft computing techniques are used [18]. So given the limited data set for maintainability metrics in the leading research, the first batch were completely removed.

Fuzzy systems, neural networks, Case-Based Reasoning (CBR) and Bayesian networks are some models based on Hierarchical Dimensional Assessment Model. further, the introduced methods, by ingratiating the desired modeling attributes namely Output Explanation ability, being suitable for small data sets, adjustment to new data, visibility of Reasoning process, being suitable for complex models, together with known facts from experts as well as by emphasizing compatibility with service-oriented architecture characteristics was evaluated and consequently in the end, fuzzy system was selected as an appropriate method [16].

As proposed evaluation structure includes two kinds of predictor or independent variables namely design phase metrics and operation phase metrics, so each of them needs a separate fuzzy system. A Discrete collection of real values from structural properties metrics namely coupling, cohesion and granularity form the fuzzy systems inputs which belong to service design phase. Also, real values or scores from selected processes maturity level evaluation include incident management, problem management, change management, configuration management; release management and availability management are fuzzy system inputs that belong to the operation phase metrics.

According to the type of problem and real value of the evaluation model inputs, the most suitable type of fuzzy system to use in this model is fuzzy system with fuzzier and defuzzier. In this type of fuzzy system, a fuzzier transforms real value of inputs into a fuzzy set as well as a defuzzier transforms fuzzy value output into a real value. This type of fuzzy system, in addition to the mentioned parts namely fuzzier and defuzzier, it has two other parts of logic rules and logic engine. TMF membership function, Centroid Average (CA) defuzzier, Mamdani logic engine is selected for the construction of the metric evaluation method.

The only issue remaining with reference to the introduced evaluation fuzzy system is the creation of logic rules and their related approvals by the experts in that field. Measuring maintainability, relations and the effects of dependent and independent variables in the service design and operation phase which have been identified in the previous section, defined in the form of fuzzy rules and through a questionnaire was validated and approved by service-oriented experts.

III. PROPOSED MODEL FOR SERVICE-ORIENTED ARCHITECTURAL MAINTAINABILITY EVALUATION

To design the model, in the previous sections the proposed solutions to solve each one of the maintainability evaluation requirements were introduce in two service design and operation phase. In this section the proposed model is offered according to the previous concepts.

A. Overall conceptual model of maintainability evaluation

The proposed model consists of five sections: input, analysis, measurement, decision making and output. In *"Fig. 1"* Components of model and their relations are presented.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

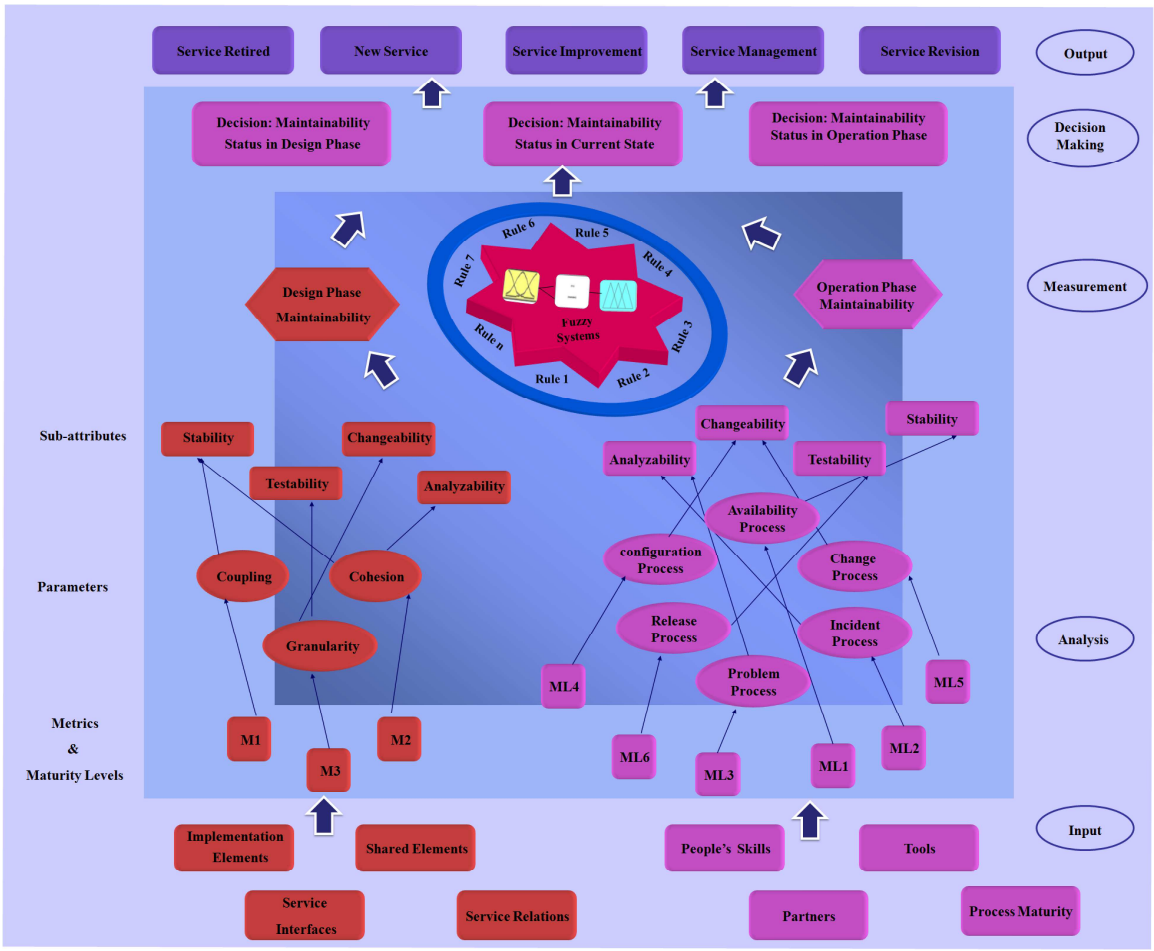


Figure 1. Components of model and their relations

Input

The inputs of design section of maintainability evaluation model include all types of service-oriented architecture relationship. In this part, software services were derived from business services in a form of atomic or compound services being analyzed. The relevant information of service component including Implementation elements, service interface and the relationship between them are obtained through an interview with the service owner or by surveying the technical documentation design and handed over to the analysis section. Additionally, the information is received from the operation section inputs from organizational experts or service owners through a questionnaire.

Analysis

This section of the proposed model includes the relationships between dependent and independent variables in the design and operation phase. In other words, this part consist of the relationship between maintainability variables with coupling, cohesion, service granularity variables and also association of the former three variables with related metrics in the design phase. Also, the rules defined between the model's different levels (sub-attributes, factors and metrics) in the design phase which have been previously approved and validated by the SOA experts is placed in the analysis section. It must be noted that similarly, information related to sub-attributes, factors, metrics and their relationship in the operation phase are also placed in the analysis section.

Measurement

This section of the model includes performing a set of rules that have been collected in the analysis section about service. By using fuzzy logic, the measurement section analyzes the collected information from the analysis section. In another word, measurement section is a collection of math functions and formulas which are based on collected information from the previous section. This part evaluates the maintainability based on the fuzzy system in each of service design and operation phase. The operation mechanism in design section is to facilitate the assessment tool receives coupling, cohesion and granularity metrics relevant information from analysis section, next by means of defined rules begins to evaluate the maintainability. Also in operation section, scores resulting from maturity level questionnaire (OGC) is received from analysis section, and

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

then maintainability of operation phase is evaluated by using associated fuzzy rules.

Decision making

As mentioned, this model provides decision making possibility about the maintainability status after the completion of the service design phase and before their operation phase and even after the completion of the operation phase. In another words, measurement section results in design section allow a service owner or manager to adopt the necessary decisions and give a recommendation about the maintainability status of software services in the design phase. Also, at a stage when the organization's software services are or suppose to be operational, service owner or manager by using measurement section results in the operation section will have the opportunity to pass judgment about the service maintainability status in operation phase. In addition when the maintainability of software services weren't evaluated in the design phase, by using this model and utilizing the measurement section in both service design and operation, a decision about the maintainability status could be made.

Output

Maintainability evaluation model output is the different decisions about the maintainability status of software services. In other words, based on the model's decision making section, in the design section, the service manager or owner will be able to take the essential action regarding the continuance of the service production, stop or making adjustment in the completed designs. Also in the operational section based on decision making section results, the service manager or owner will have the opportunity to plan and take the necessary action regarding improvements in processes, people, product and provider in support area of the service management. Also regarding the live software service, the mentioned model will provide ability for the evaluation of the maintainability status in the service design and operation phase of software service for the service manager or owner.

IV. CONCLUSION

In this article, by considering various factors in total service lifecycle affecting the service maintainability, a practical and comprehensive service maintainability evaluation model in service-oriented architecture were proposed. This model includes five sections: input, analysis, measurement, decision making and output. The relationship between independent variables (cohesion, coupling and granularity) in the service design phase with the maintainability dependent variable was determined through a questionnaire completed by the service-oriented architecture experts. Also, relationship between dependent variables meaning six process factors (incident management, problem management, change management, configuration management, release management and availability management) with the maintainability dependent variable was identified through the completion of a questionnaire.

Further, based on analysis of the collected information, fuzzy rules were define and used to evaluate the maintainability in the service lifecycle. This model provides the possibility to judge and make decisions about the software service maintainability status in every step of the service lifecycle. So based on these decisions, the owner and manager will be able to take control effort or make the necessary corrections in the fastest possible time.

REFERENCES

[1] Bingu Shim, Siho Choue, Suntae Kim, Sooyong Park, "A Design Quality Model for Service-Oriented Architecture," 15th Asia-Pacific Software Engineering Conference, 2008.

[2] Mikhail Perepletchikov, Caspar Ryan, and Zahir Tari, "The Impact of Software Development Strategies on Project and Structural Software Attributes in SOA," ARC (Australian Research Council), under Linkage scheme no. LP0455234.

[3] Mikhail Perepletchikov and Caspar Ryan, "A Controlled Experiment for Evaluating the Impact of Coupling on the Maintainability of Service-Oriented Software," IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, 2009.

[4] Mikhail Perepletchikov, Caspar Ryan, Keith Frampton, and Zahir Tari, "Coupling Metrics for Predicting Maintainability in Service-Oriented Designs," Proceedings of the 2007 Australian Software Engineering Conference (ASWEC'07), 2007.

[5] Mikhail Perepletchikov, Caspar Ryan, and Zahir Tari, "The Impact of Service Cohesion on the Analysability of Service-Oriented Software," IEEE TRANSACTIONS ON SERVICES COMPUTING, 2009.

[6] Mikhail Perepletchikov, Caspar Ryan, and Keith Frampton, "Cohesion Metrics for Predicting Maintainability of Service-Oriented Software," Seventh International Conference on Quality Software, 2007.

[7] Mehwish Riaz, Emilia Mendes, Ewan Tempero, "A Systematic Review of Software Maintainability Prediction and Metrics," Third International Symposium on Empirical Software Engineering and Measurement, 2009.

[8] Renuka Sindhgatta, Bikram Sengupta and Karthikeyan Ponnalagu, "Measuring the Quality of Service Oriented Design," Lecture Notes in Computer Science, 2009, Volume 5900/2009, 485-499, DOI: 10.1007/978-3-642-10383-4_36.

[9] Wang Xiao-jun, "Metrics for Evaluating Coupling and Service Granularity in Service Oriented Architecture," IEEE supported by National Key Technology R&D Program (No.2007BAH17B04), and Research Climbing Project of NJUPT(No.NY207062), 2009.

[10] Pham Thi Quynh, Huynh Quyet Thang, "Dynamic Coupling Metrics for Service – Oriented Software," International Journal of Computer Science and Engineering 3:1, 2009.

[11] Qingqing Zhang, Xinke Li, "Complexity Metrics for Service-Oriented Systems," Second International Symposium on Knowledge Acquisition and Modeling, 2009.

[12] Helge Hofmeister and Guido Wirtz, "Supporting Service-Oriented Design with Metrics," 12th International IEEE Enterprise Distributed Object Computing Conference, 2008.

[13] Sharma, A., Grover, P.S., Kumar, R., "Predicting Maintainability of Component-Based Systems by Using Fuzzy Logic," IC3(2009) 581-591.

[14] IEEE Std. 610.12-1990, "Standard Glossary of Software Engineering Terminology," IEEE Computer Society Press, Los Alamitos, CA, 1993.

[15] ITIL definition: Maintainability (ITILv3): http://www.knowledgetransfer.net/dictionary/ITIL/en/Maintainability.htm

[16] Andrew R. Gray and Stephen G. MacDonell, "A Comparison of Techniques for Developing Predictive Models of Software Metrics," Information and Software Technology 39: 425-437, 1997.

[17] Emad Ghosheh , Jihad Qaddour , Matthew Kuofie and Sue Black, "A comparative analysis of maintainability approaches for web," IEEE International Conference on Computer Systems and Applications (2006) pp.1155-1158.

[18] Deepak Gupta, Vinay Kr.Goyal, Harish Mittal, "Comparative Study of Soft Computing Techniques for Software Quality Model," International Journal of Software Engineering Research & Practices Vol.1, Issue 1, Jan, 2011.

[19] Taixi Xu, Kai Qian, Xi He, "Service Oriented Dynamic Decoupling Metrics," International Conference on Semantic Web and Web Services (SWWS'06) June 26-29, 2006 WORLDCOMP'06, Las Vegas, USA, 2006.

[20] Kai Qian, Jigang Liu, Frank Tsui, "Decoupling Metrics for Services Composition," Proceedings of the 5th IEEE/ACIS International Conference on Computer and Information Science and 1st IEEE/ACIS, 2006.

[21] A. Karahasanovic, A. K. Levine, and R. C. Thomas, "Comprehension strategies and difficulties in maintaining object-oriented systems: An explorative study," Journal of Systems and Software, vol. 80 (9), pp. 1541-1559, 2007.

[22] M. Garcia and J. Alvarez, "Maintainability as a key factor in maintenance productivity: A case study," International Conference on Software Maintenance (ICSM), Washington, USA, pp. 87-93, 1996.

[23] S. Muthanna, K. Kontogiannis, K. Ponnambalam, et al., "A maintainability model for industrial software systems using design level metrics," Seventh Working Conference on Reverse Engineering, Brisbane, Australia, p. 248, 2000.

[24] A. Takang and P. Grubb: Software Maintenance, "Concepts and Practice," London: Thompson Computer Press,1996.

[25] D. P. Darcy, C. F. Kemerer, S. A. Slaughter, et al., "The structural complexity of software: an experimental test," IEEE Transactions on Software Engineering, vol. 31 (11), pp. 982 – 995, 2005.

[26] J. Eder, G. Kappel, and M. Schrefl, "Coupling and Cohesion in Object-Oriented Systems," ACM Conference on Information and Knowledge Management, 1992.

[27] "ISO/IEC 9126-1:2001 Software Engineering: Product quality – Quality model," International Standards Organisation, Geneva, 2001.

[28] ITIL Service Management Self Assessment, http://www.itsmf.com/trans/sa.asp.

[29] N.Zhou, T.Zhu, and H. Wang, "Evaluating Service Identification With Design Metric on Business Process Decomposition," October, IEEE International Conference on Services Computing, 2009.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The SVM Based Interactive tool for Predicting Phishing Websites

Santhana Lakshmi V
Research Scholar,
PSGR Krishnammal College for Women
Coimbatore,Tamilnadu.
sanlakmphil@gmail.com

Vijaya MS
Associate Professor,Department of Computer Science,
GRG School of Applied Computer Technology,
Coimbatore,Tamilnadu.
msvijaya@grgsact.com

***Abstract*— Phishing is a form of social engineering in which attackers endeavor to fraudulently retrieve the legitimate user's confidential or sensitive credentials by imitating electronic communications from a trustworthy or public organization in an automated fashion. Such communications are done through email or deceitful website that in turn collects the credentials without the knowledge of the users. Phishing website is a mock website whose look and feel is almost identical to the legitimate website. So internet users expose their data expecting that these websites come from trusted financial institutions. Several antiphishing methods have been introduced to prevent people from becoming a victim to these types of phishing attacks. Regardless of the efforts taken, the phishing attacks are not alleviated. Hence it is more essential to detect the phishing websites in order to preserve the valuable data. This paper demonstrates the modeling of phishing website detection problem as binary classification task and provides convenient solution based on support vector machine, a pattern classification algorithm. The phishing website detection model is generated by learning the features that have been extracted from phishing and legitimate websites. A third party service called 'blacklist' is used as one of the feature that helps to envisage the phishing website effectively. Various experiments have been carried out and the performance analysis shows that the SVM based model outperforms well.**

Keywords- Antiphishing, Blacklist, Classification, Machine Learning, Phishing, Prediction

INTRODUCTION

Phishing is a novel crossbreed of computational intelligence and technical attacks designed to elicit personal information from the user. The collected information is then used for a number of flagitious deeds including fraud, identity theft and corporate espionage. The growing frequency and success of these attacks led a number of researchers and corporations to take the problem seriously. Various methodologies are adopted at present to identify phishing websites. Maher Aburous et, al. proposes an approach for intelligent phishing detection using fuzzy data mining. Two criteria are taken into account. URL - domain identity and Security-Encryption [1]. Ram basnet et al. adopts machine learning approach for detecting phishing attacks. Biased support vector machine and Neural Network are used for the efficient prediction of phishing websites [2]. Ying Pan and Xuhus Ding used anomalies that exist in the web pages to detect the mock website and support vector machine is used as a page classifier [3]. Anh Le, Athina Markopoulou, University of California used lexical features of the URL to predict the phishing website. The algorithms used for prediction includes support vector machine, Online perceptron, Confidence-Weighted and Adaptive Regularization of weights [4]. Troy Ronda have designed an anti phishing tool that does not rely completely on automation to detect phishing. Instead it relies on user input and external repositories of information [5].

In this paper, the detection of phishing websites is modelled as binary classification task and a powerful machine-learning based pattern classification algorithm namely support vector machine is employed for implementing the model. Training the features of phishing and legitimate websites helps to create the learned model.

Feature extraction method presented here is similar to the one presented in [3] [6] [7] and [8]. The features such as foreign anchor, nil Anchor, IP address, dots in page address, dots in URL, slash in page address, slash in URL, foreign Anchor in identity set, Using @ Symbol, server form handler (SFH), foreign request, foreign request URL in identity set, cookie, SSL certificate, search engine, 'Whois' lookup, used in their work are taken into account in this work. But some of the features such as hidden fields and age of the domain are omitted since they do not contribute much for predicting the phishing website.

Hidden field is similar to the text box used in HTML except that the hidden box and the text within the box will not be visible as in the case of textbox. Legitimate websites also use hidden fields to pass the user's information from one form to another form without forcing the users to re-type over and over again. So presence of hidden field in a webpage cannot be considered as a sign of being a phishing website.

Similarly age of the domain specifies the life time of the websites in the web. Details regarding the life time of a website can be extracted from the 'Whois' database which contains the registration information of all the users. Legitimate websites

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

have long life when compared to phishing websites. But this feature cannot be considered to recognize the phishing websites since the phishing web pages that are hosted on the compromised web server also contains long life. The article [9] provides empirical evidence according to which 75.8% of the phishing sites that are analyzed (2486 sites) were hosted on compromised web servers to which the phishers obtained access through google hacking techniques.

This research work makes use of certain features that were not taken into consideration in [6]. They are 'Whois' look up and server form handler. 'Whois' is a request response protocol used to fetch the registered customer details from the database. The database contains the information such as primary domain name, registrar, registration date, expiry date of a registered website. The legitimate website owners are the registered users of 'whois' database. The details of phishing websites will not be available in 'whois' database. So the existence of a websites' details in 'whois' database is an evidence for being legitimate. So it is essential to use this feature for identifying the phishing websites.

Similarly in case of server form handler, HTML forms that include textbox, checkbox, buttons etc are used to pass data given by the user to a server. Action is a form handler and is one of the attributes of form tag, which specifies the URL to which the data should be transferred. In the case of phishing websites, it specifies the domain name, which embezzles the credential data of the user. Even though some legitimate websites use third party service and hence may contain foreign domain, it is not the case for all the websites. So it is cardinal to check the handler of the form. If the handler of a form points to a foreign domain it is considered to be a phishing website. Instead if the handler of a website refers to the same domain, then the website is considered as legitimate. Thus these two features are very much essential and hope to contribute more in classifying the website.

The research work described here also seeks the usage of third party service named 'Blacklist' for predicting the website accurately. Blacklist contains the list of phishing and suspected websites. The page URL is checked against 'Blacklist' to verify whether the URL is present in the blacklist.

The process of identity extraction and feature extraction are described in the following section and the various experiments carried out to discover the performance of the models are demonstrated in the rest of this paper.

I. PROPOSED PHISHING WEBSITE DETECTION MODEL

Phishing websites are replica of the legitimate websites. A website can be mirrored by downloading and using the source code used for designing the website. Before acquiring these websites, their source code is captured and parsed for DOM objects. Identities of these websites are extracted from the DOM objects. The main phase of phishing website prediction is identity extraction and feature extraction. Essential features that contribute to the detection of the category of the websites, whether phishing or legitimate are extracted from the URL and source code for envisaging the phishing websites accurately. The training dataset with instances pertaining to legitimate and phishing websites is developed and used for learning the model. The trained model is then used for predicting unseen instance of a website. The architecture of the system is shown in figure Figure 1.

Figure 1. System Architecture

A. 2.1 Identity Extraction

Identity of a web page is a set of words that uniquely determines the proprietorship of the website. Identity extraction should be accurate for the successful prediction of phishing website. In spite of phishing artist creating the replica of legitimate website, there are some identity relevant features which cannot be exploited. The change in these features affects the similarity of the website. This paper employs anchor tag for identity extraction. Anchor tag is used to find the identity of a web page accurately. The value of the href attribute of anchor tag has high probability of being an identity of a web page. Features extracted in identity extraction phase include META Title, META Description, META Keyword, and HREF of <a> tag.

META Tag

The <Meta> tag provides metadata about the HTML document. Metadata will not be displayed on the page, but will be machine parsable. Meta elements are typically used to specify page description, keywords, author of the document, last modified and other metadata. The <Meta> tag always goes inside the head element. The metadata is used by the browsers to display the content or to reload the page, search engines, or other web services.

META Description Tag

The Meta description tag is a snippet of HTML code that comes inside the <Head> </Head> section of a Web page. It is usually placed after the Title tag and before the Meta keywords tag, although the order is not important. The proper syntax for this HTML tag is

"<META NAME="Description" CONTENT="Your descriptive sentence or two goes here.">"

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The identity relevant object is the value of the content attribute. The value of the content attribute gives brief description about the webpage. There is a greater possibility for the domain name to appear in this place.

META Keyword Tag

The META Keyword Tag is used to list the keywords and keyword phrases that were targeted for that specific page.

<META NAME="keywords" content="META Keywords Tag, Metadata Elements, Indexing, Search Engines, Meta Data Elements">

The value of the content attribute provides keywords related to the web page.

HREF

The href attribute of the <a> tag indicates the destination of a link. The value of the href attribute is a URL to which the user has to be directed. When the hyperlinked text is selected, users should be directed to the concerned web page. Phishers do change this value. Since any change in the appearance of the webpage may reveal the users that the websites is forged. So the domain name in the URL has high probability to be the identity of the website.

Once the identity relevant features are extracted, they are converted into individual terms by removing the stop words such as http, www, in, com, etc., and by removing the words with length less than three. Since the identity of a website is not expected to be very small. Tf-idf weight is evaluated for each of the keywords. The first five keywords that have high tf-idf value are selected for identity set. tf-idf value is calculated using the following formula.

$$tf_{ij} = \sqrt{\frac{n_{ij}}{\sum_k n_{kj}}} \quad (1)$$

where n_{kj} is the number of occurrence of t_i in document d_j and $\sum_k n_{kj}$ is the number of all terms in document d_j.

$$idf_i = ln\left(\frac{|D|}{|\{d_j : t_i \sum d_j\}| + 1}\right) \quad (2)$$

Where |D| is the total number of documents in a dataset, and $\{|dj:t_i \in dj\}|$ is the number of documents where term ti appears. To find the document frequency of a term, WebAsCorpus is used. It is a readymade frequency list. The list contains words and the number of documents in which the words appear. The total number of documents in which the term appears is the term that has the highest frequency. The highest frequency term is assumed to be present in all the documents.

The tf-idf weight is calculated using the following formula

$$tf - idf_{ij} = tf_{ij} \cdot idf_i \quad (3)$$

The keywords that have high tf-idf weight are considered to have greater probability of being the web page identity.

II FEATURE EXTRACTION AND GENERATION

Feature extraction plays an important role in improving the classification effectiveness and computational efficiency. Distinctive features that assist to predict the phishing websites accurately are extracted from the corresponding URL and source code. In a HTML source code there are many characteristics and features that can distinguish the original website from the forged websites. A set of 17 features are extracted for each website to form a feature vector and are explained below.

- Foreign Anchor

An anchor tag contains href attribute. The value of the href attribute is a URL to which the page is linked with. If the domain name in the URL is not similar to the domain in page URL then it is considered as foreign anchor. Presence of too many foreign anchor is a sign of phishing website. So all the href values of <a> tags used in the web page are examined. And they are checked for foreign anchor. If the number of foreign domain exceeds, then the feature F_1 is assigned to -1. Instead if the webpage contains minimum number of foreign anchor, the value of F_1 is 1.

- Nil Anchor

Nil anchors denote that the page is linked with no page. The value of the href attribute of <a> tag will be null. The values that denote nil anchor are about: blank, JavaScript::, JavaScript: void(0), #. If these values exist then the feature F_2 is assigned the value of –1.Instead the value of F_2 is assigned as 1.

- IP Address

The main aim of phishers is to gain lot of money with no investment and they will not spend money to buy domain names for their fake website. Most phishing websites contain IP address as their domain name. If the domain name in the page address is an IP Address then the value of the feature F_3 is –1 else the value of F_3 is 1.

- Dots in Page Address

The page address should not contain more number of dots. If it contains more number of dots then it is the sign of phishing URL. If the page address contains more than five dots then the value of the feature F_4 is -1 or else the value of F_4 is 1.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Dots in URL

This feature is similar to feature F_4.But here the condition is applied to all the urls including href of <a> tag, src of image tag etc., All the url's are extracted and checked. If the URL contains more than five dots then the value of the feature vector F_5 is -1 or else the value of F_5 is 1.

- Slash in page address

The page address should not contain more number of slashes. If the page url contains more than five slashes then the url is considered to be a phishing url and the value of F_6 is assigned as –1. If the page address contains less than 5 slashes, the value of F_6 is 1.

- Slash in URL

This feature is similar to feature F_6. But the condition is checked against all the urls used in the web page. If the urls collected have more than five slashes, the feature F_7 is assigned –1. Instead the value of F_7 is 1.

- Foreign Anchor in Identity Set

Phishing artist makes slight changes to the page URL to make it believe as legitimate URL. But changes cannot be made to all the urls used in the source code. So the urls used in the source code will be similar to the legitimate website. If the website is legitimate, then both the url and the page address will be similar and it will be present in the identity set. But while considering phishing website, the domain of the URL and the page address will not be identical and domain name will not be present in the identity set. If the anchor is not a foreign anchor and is present in identity set then the value of F_8 is 1.If the anchor is a foreign anchor but present in the identity set then also the value of F_8 is 1.If the anchor is a foreign anchor and is not present in the identity set then the value of F_8 is –1.

- Using @ Symbol

Page URL that are longer than normal, contain the @ symbol. It indicates that the all text before @ is comment. So the page url should not contain @ symbol. If the page URL contains @ symbol, the value of F_9 Is –1 otherwise the value is assigned as +1.

- Server Form Handler (SFH)

Forms are used to pass data to a server. Action is one of the attributes of form tag, which specifies the url to which the data should be transferred. In the case of phishing website, it specifies the domain name, which embezzles the credential data of the user. Even though some legitimate websites use third party service and hence contain foreign domain, it is not the case for all the websites. It is cardinal to check the value of the action attribute. The value of the feature F_{10} is –1, if the following conditions hold. 1) The value of the action attribute of form tag comprise foreign domain, 2) value is empty, 3) value is #, 4) Value is void. If the value of the action attribute is its own domain then, F_{10}= 1.

- Foreign Request

Websites request images, scripts, CSS files from other place. Phishing websites to imitate the legitimate website request these objects from the same page as legitimate one. The domain name used for requesting will not be similar to page URL. Request urls are collected from the src attribute of the tags <img> and <script>, background attribute of body tag, href attribute of link tag and code base attribute of object and applet tag. If the domain in these urls is foreign domain then the value of F_{11} is –1 or else the value is 1.

- Foreign request url in Identity set

If the website is legitimate, the page url and url used for requesting the objects such as images, scripts etc., should be similar and the domain name should be present in the identity set. The entire request URL in the page is checked for the existence in identity set. If they exist the value of F_{12} is 1.If they does not exist in the identity set the value of F_{12} is –1.

- Cookie

Web cookie is used for an original website to send state information to a user's browser and for the browser to return the state information to the website. In simple it is used to store information. The domain attribute of cookie holds the server domain, which set the cookies. It will be a foreign domain for phishing website. If the value of the domain attribute of cookie is a foreign domain then F_{13} is -1 otherwise F_{13} is 1. Some websites do not use cookies. If no cookies found then F_{13} is 2.

- SSL Certificate

SSL is an acronym of secure socket layer. SSL creates an encrypted connection between the web server and the user's web browser allowing for private information to be transmitted without the problems of eavesdropping, data tampering or message forgery. To enable SSL on a website, it is required to get an SSL Certificate that identifies the user and install it on the server. All legitimate websites will have SSL certificate. But phishing websites do not have SSL certificate. The feature corresponding to SSL certificate is extracted by providing the page address. If the SSL certificate exists for the website then the value of the feature F_{13} is 1.If there is no SSL certificate then the value of F_{13} is -1.

- Search Engine

If the legitimate website's URL is given as a query to search engine, then the first results produced should be related to the concerned website. If the page URL is fake, the results will not be related to the concerned website. If the first 5 results from the search engine is similar to the page URL then the value of F14 is 1.Otherwise the value of F_{14} is assigned as -1.

- 'Whois' Lookup

'Whois' is a request response protocol is used to fetch the registered customer details from the database. The database contains the information about the registered users such as registration date, duration, expiry date etc. The legitimate site owners are the registered users of 'whois' database. The details of phishing website will not be available in 'whois' database. 'Whois' database is checked for the existence of the data pertaining to a particular website. If exists then the value of F_{16} is 1 or F_{16} is assigned as –1.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Blacklist

Blacklist contains list of suspected websites. It is a third party service. The page URL is checked against the blacklist. If the page URL is present in the blacklist it is considered to be a phishing website. If the page URL exist in the blacklist then the value of F_{17} is –1 otherwise the value is 1.

Thus a group of 17 features describing the characteristics of a website are extracted from the HTML source code and the url of a website by developing PHP code. The feature vectors are generated for all the websites and the training dataset is generated.

III. SUPPORT VECTOR MACHINE

Support vector machine represents a new approach to supervised pattern classification, which has been successfully applied to a wide range of pattern recognition problems. It is a new generation learning system based on recent advances in statistical learning theory [10]. SVM as supervised machine learning technology is attractive because it has an extremely well developed learning theory, statistical learning theory. SVM is based on strong mathematical foundations and results in simple yet very powerful algorithms. SVM has a number of interesting properties, including the solution of Quadratic Programming problem is globally optimized, effective avoidance of over fitting, the ability to handle large feature spaces, can identify a small subset of informative points called SV and so on.

The SVM approach is superior in all practical applications and showing high performances. For the last couple of years, support vector machines have been successfully applied to a wide range of pattern recognition problems such as text categorization, image classification, face recognition, hand written character recognition, speech recognition, biosequence analysis, biological data mining, Detecting Steganography in digital images, Stock Forecast, Intrusion Detection and so on. In these cases the performance of SVM is significantly better than that of traditional machine learning approaches, including neural networks.

Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a new data point will be in. In the case of support vector machines, a data point is viewed as a *p*-dimensional vector of a list of *p* numbers, and one wants to know whether one can separate such points with a p − 1-dimensional hyper plane. This is called a linear classifier. There are many hyper planes that might classify the data. The maximum separation of margin between the two classes is usually desired [11]. So choose the hyper plane so that the distance from it to the nearest data point on each side is maximized. If such a hyper plane exists, it is clearly of interest and is known as the maximum-margin hyper plane and such a linear classifier is known as a maximum margin classifier.It is the simplest models SVM based maximal margin. If w is weight vector realizing functional margin 1 on the positive point X^+ and on the negative point X^- , then the two planes parallel to the hyper plane which passes through one or more points called bounding hyper planes are given by

$$W^T X - \gamma = 1$$
$$W^T X - \gamma = -1 \quad (4)$$

The margin between the optimal hyper plane and the bounding plane is 1/||w||, and so the distance between the bounding hyper planes is 2/||w||. Distance of the bounding plane $w^T x - \gamma = 1$ from the origin is $|-\gamma + 1|/||w||$ and the distance of the bounding plane $w^T x - \gamma = -1$ from the origin is $|-\gamma - 1|/||w||$.

The points falling on the bounding planes are called support vectors and these points play crucial role in the theory. The data points x belonging to two classes A+ and A- are classified based on the condition.

$$W^T X_i - \gamma \geq 1 \text{ for all } X_i \in A^+$$
$$W^T X_i - \gamma \leq -1 \text{ for all } X_i \in A^- \quad (5)$$

These inequality constraints can be combined to give

$$D_{ii}(W^T X_i - \gamma) \geq 1 \text{ for all } X_i \quad (6)$$

Where $D_{ii} = 1$ for A^+ and $D_{ii}=-1$ for A^- . The

learning problem is hence to find an optimal hyper plane $<w, \gamma>$, $w^T x - \gamma = 0$ which separates A^+ from A^- by maximizing the distance between the bounding hyper planes. Then the learning problem is formulated as an optimization problem as below

$$Minimize = \frac{1}{2}|W|^2 \quad (7)$$
$$Subject\ to\ D_{ii}(W^T X_i - \gamma) \geq 1 \quad i = 1,2,\ldots\ldots,l.$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. EXPERIMENT AND RESULTS

The phishing website detection model is generated by implementing SVM using SVM [light]. It is an implementation of Vapnik's Support Vector Machine for the problem of pattern recognition, for the problem of regression, and for the problem of learning a ranking function. The dataset used for learning are collected from PHISHTANK [12]. It is an archive consisting of collection of phishing websites. The dataset with 150 phishing websites and 150 legitimate websites are developed for implementation. The features describing the properties of websites are extracted and the size of each feature vector is 17. The feature vector corresponding to phishing website is assigned a class label -1 and +1 is assigned for legitimate website.

The experiment and data analysis is also carried out using other classification algorithms such as multilayer perceptron, decision tree Induction and naïve Bayes in WEKA environment for which the same training dataset is employed. The Weka Open source, portable, GUI-based workbench is a collection of state-of-the-art machine learning algorithms and data pre-processing tools. For Weka the class label is assigned as 'L' that denotes legitimate websites and 'P' for phishing websites

A. *Classification Using SVM [light]*

The dataset is trained with linear, polynomial and RBF kernel with different parameter settings for C- regularization parameter. In case of polynomial and RBF kernels, the default settings for d and gamma are used. The performance of the trained models is evaluated using 10-fold cross validation for its predictive accuracy. Predictive accuracy is used as a performance measure for phishing website prediction. The prediction accuracy is measured as the ratio of number of correctly classified instances in the test dataset and the total number of test cases. The performances of the linear and non-linear SVM classifiers are evaluated based on the two criteria, the prediction accuracy and the training time.

Regularization parameter C is assigned different values in the range of 0.5 to 10 and found that the model performs better and reaches a stable state for the value C = 10**.** The performance of the classifiers are summarized in Table IV and shown in Fig.2 and Fig.3.

The result of the classification model based on SVM with linear kernel is shown Table I

TABLE 1 LINEAR KERNEL

Linear SVM	C=0.5	C=1	C=10
Accuracy(%)	91.66	95	92.335
Time(S)	0.02	0.02	0.03

The results of the classification model based on SVM with polynomial kernel and with parameters d and C are shown in Table II.

TABLE 2. POLYNOMIAL KERNEL

d	C=0.5		C=1		C=10	
	1	*2*	*1*	*2*	*1*	*2*
Accuracy (%)	97.9	98.2	90	90.1	96.3	96.08
Time	0.1	0.3	0.1	0.8	0.9	0.2

The predictive accuracy of the non-linear support vector machine with the parameter gamma (g) of RBF kernel and the regularization parameter C is shown in Table III.

TABLE 3. RBF KERNEL

g	C=0.5		C=1		C=10	
	1	*2*	*1*	*2*	*1*	*2*
Accuracy(%)	99.2	99.1	98.6	98.3	97.4	97.1
Time	0.1	0.1	0.2	0.2	0.1	0.1

The average and comparative performance of the SVM based classification model in terms of predictive accuracy and training time is given in Table IV and shown in Fig.2 and Fig.3

TABLE 4. AVERAGE PERFORMANCE OF THREE MODELS

Kernels	Accuracy	Time taken to build model(s)
Linear	92.99	0.02
Polynomial	94.76	0.4
RBF	98.28	0.13

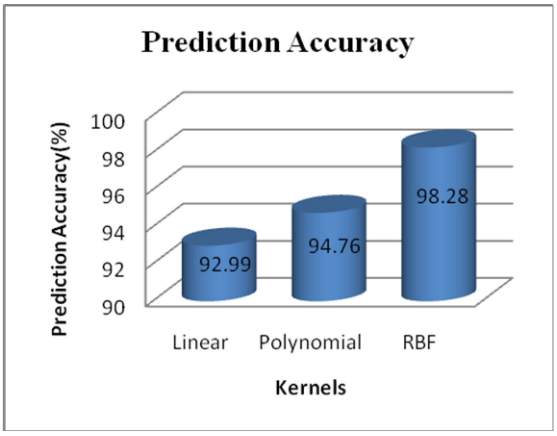


Figure 7. Prediction Accuracy

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

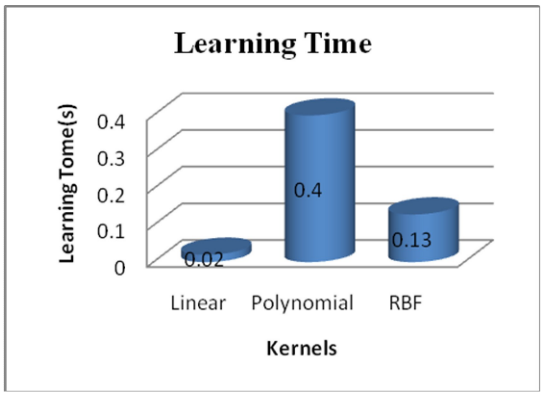


Figure 8. Prediction Accuracy

From the above comparative analysis the predictive accuracy shown by SVM with RBF kernel is higher than the linear and polynomial SVM. The time taken to build the model using SVM with polynomial kernel is more, than linear and RBF kernel.

B. Classification Using Weka

The classification algorithms, multi Layer perceptron, decision tree induction and naïve bayes are implemented and trained using WEKA. The Weka, Open Source, Portable, GUI-based workbench is a collection of state-of-the-art machine learning algorithms and data pre processing tools [13] [20]. The robustness of the classifiers is evaluated using 10 fold cross validation. Predictive accuracy is used as a primary performance measure for predicting the phishing website. The prediction accuracy is measured as the ratio of number of correctly classified instances in the test dataset and the total number of test cases. The performances of the trained models are evaluated based on the two criteria, the prediction accuracy and the training time. The prediction accuracy of the models is compared.

The 10-fold cross validation results of the three classifiers multilayer perceptron, decision tree induction and naive bayes are summarized in Table V and Table VI and the performance of the models is illustrated in figures Fig 4 and Fig 5.

TABLE-V PERFORMANCE COMPARISON OF CLASSIFIERS

Evaluation Criteria	Classifiers		
	MLP	***DTI***	***NB***
Time taken to build model (Secs)	1.24	0.02	0
Correctly Classified instances	282	280	281
Incorrectly Classified instances	18	20	19
Prediction accuracy (%)	94	93.333	93.667

TABLE-VI COMPARISON OF ESTIMATES

Evaluation Criteria	Classifiers		
	MLP	***DT***	***NB***
Kappa statistic	0.88	0.8667	0.8733
Mean Absolute Error	0.074	0.1004	0.0827
Root Mean Squared error	0.2201	0.2438	0.2157
Relative absolute error	14.7978	20.0845	16.5423
Root relative square error	44.0296	48.7633	43.1423

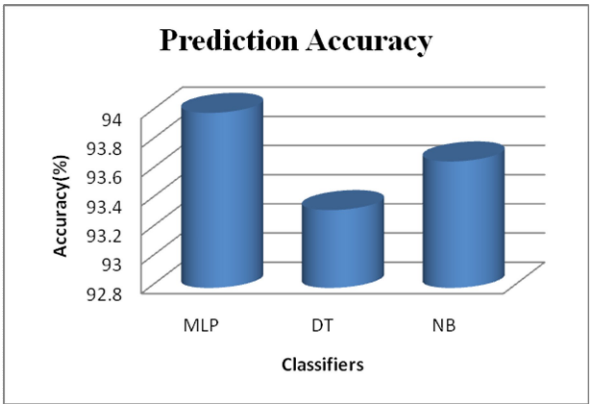


Figure 9. Prediction Accuracy

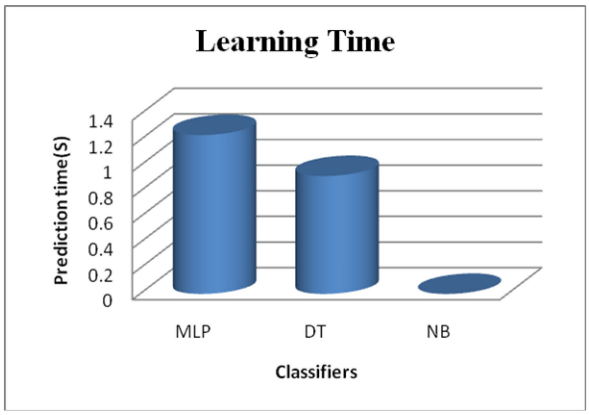


Figure 10. Learning Time

The time taken to build the model and the prediction accuracy is high in the case of naïve bayes, when compared to other two algorithms. As far as the phishing website prediction system is concerned, predictive accuracy plays major role than learning time in predicting whether the given website is phishing or legitimate.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

V. PHISHING WEBSITE PREDICTION TOOL

Phishing Website prediction tool is designed and classification algorithms are implemented using PHP. It is a widely-used general-purpose scripting language that is especially suited for Web development and can be embedded into HTML. In a HTML source code there are many characteristics and features that can distinguish the original website from the forged websites. The process of extracting those characteristics from a source code is called screen scraping. Screen Scraping involves scraping the source code of a web page, getting it into a string, and then parsing the required parts. Identity extraction and feature extraction are performed through screen scraping the source code. Feature vectors are generated from the extracted features.

Then feature vectors are trained with SVM to generate a predictive model using which the category of new website is discovered. Screenshots of the phishing website prediction tool are shown in Figure 2,Figure 3…Figure 7

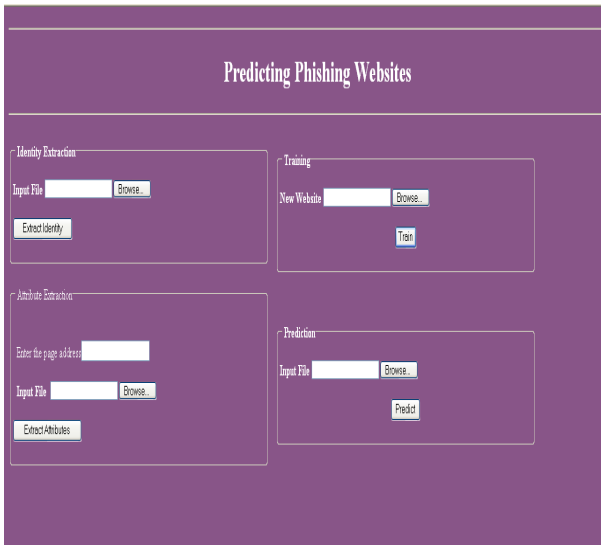


Figure 2. Phishing website prediction tool

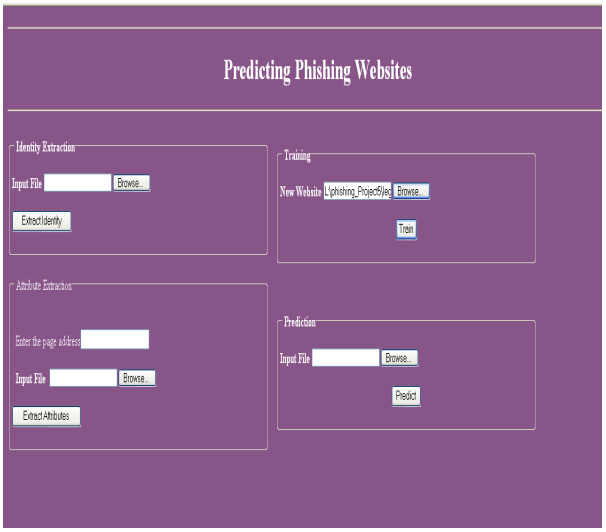


Figure 3. Training file selection

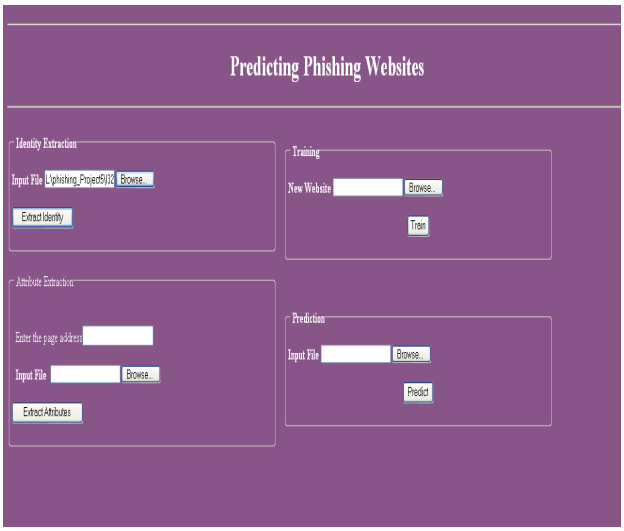


Figure 4. Identity extraction

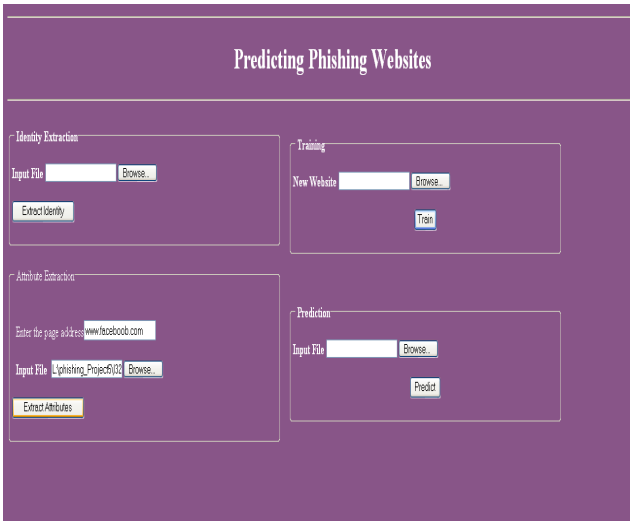


Figure 5. Feature extraction

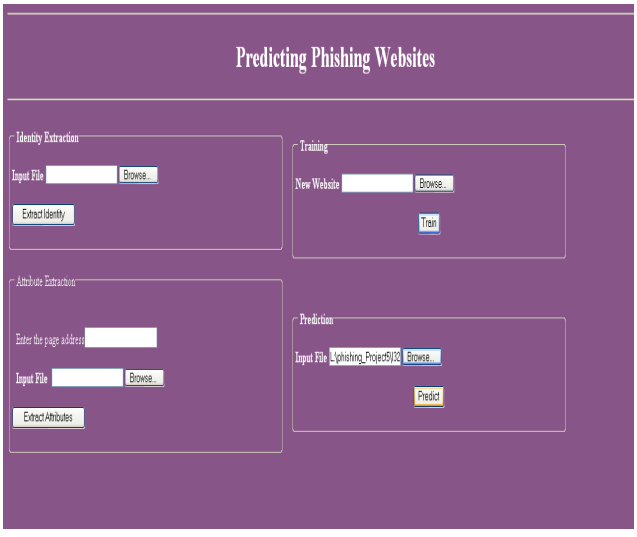


Figure 6. Testing

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 7. Prediction result

V. CONCLUSION

This paper demonstrates the modeling of phishing website detection problem as classification task and the prediction problem is solved using the supervised learning approach. The supervised classification techniques such as support vector machine, naïve bayes classifier, decision tree classifier, and multiplayer perceptron are used for training the prediction model. Features are extracted from a set of 300 URL and the corresponding HTML source code of phishing and legitimate websites. Training dataset has been prepared in order to facilitate training and implementation. The performance of the models has been evaluated based on two performance criteria, predictive accuracy and ease of learning using 10-fold cross validation. The outcome of the experiments indicates that the support vector machine with RBF kernel predicts the phishing websites more accurately while comparing to other models. It is hoped that more interesting results will follow on further exploration of data.

References

[1] Fadi Thabath, Keshav Dahal, Maher Aburrous, "Modelling Intelligent Phishing Detection System for e-Banking using Fuzzy Data Mining".

[2] Andrew H.Sung , Ram Basenet and Srinivas Mukkamala, "Detection of Phishing Attacks: A machine Learning Approach".

[3] Xuhus Ding , Ying Pan "Anomaly Based Phishing page Detection".

[4] Anh Le, Athina Markopoulou, Michalis Faloutsos "PhishDef: URL Names Say it All".

[5] Alec Wolman , Stefan Sarolu,, Troy Ronda, "iTrustpage: A User-Assisted Anti-Phishing Tool".

[6] Adi Sutanto , Pingzhi, Rong-Jian Chen, Muhammad Khurram Khan, Mingxing He, Ray-Shine Run, Shi-Jinn Horng, Jui-Lin Lai, "An efficient phishing webpage detector"

[7] Sengar PK, Vijay Kumar "Client-Side Defence against phishing with pagesafe"

[8] Lorrie Cranor , Jason Hong, Yue Zhang, "CANTINA:A Content-Based Approach to Detecting Phishing Web sites"

[9] Richard Clayton and Tyler Moore "Evil Searching: Compromise and Recompromise of Internet Hosts for Phishing"

[10] Ajay V, Loganathan R, Soman K.P, Machine Learning with SVM and other Kernel Methods, PHI, India, 2009

[11] John Shawe-Taylor, Nello Cristianini, "Support Vector Machines and other kernel- based learning methods", 2000, Cambridge University Press, UK.

[12] www.phishtank.com

[13] Eibe Frank, Ian H. Witten.2005 Data Mining – Practical Machine Learning Tools and Techniques. Elsevier Gupta GK "Introduction to Data Mining with Case Studies"

[14] Aboul Ella Hassanien , Dominik Slezak (Eds), Emilio Corchado, Javier Sedano, Jose Luis Calvo, Vaclav Snasel, "Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011.

[15] Mitchell T "Machine learning" Ed. Mc Graw-Hill International edition.

[16] Crammer K and Singer Y. "On the algorithmic implementation of Multiclass SVMs", JMLR, 2001.

[17] Lipo Wang "Support Vector Machines: theory and Applications"

[18] Burges C, , Joachims T, Schölkopf B, Smola A, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, MIT Press, Cambridge, MA, USA, 1999.

[19] Statistical Learning Theory, Vapnik VN, Wiley.J & Sons, Inc., New York, 1998.

[20] Eibe Frank, Geoffrey Holmes, lan Witten H, Len Trigg, Mark Hall Sally Jo Cunningham, "Weka: Practical Machine Learning Tools and Techniques with Java Implementations," Working Paper 99/11, Department of Computer Science, The University of Waikato, Hamilton, 1999.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Color-Base Skin Detection using Hybrid Neural Network & Genetic Algorithm for Real Times

Hamideh Zolfaghari[1]
zolfaghari_61@yahoo.com

Azam Sabbagh Nekonam [2]
aznekonam@yahoo.com

Javad Haddadnia [3]
Haddadnia@sttu.ac.ir

Department of Electronic Engineering[1,2,3]
Sabzevar Tarbeyat Moallem University
Sabzevar, Iran[1,2,3]

***Abstract*—This paper present a novel method of human skin detection base on hybrid neural network(NN) and genetic algorithm(GA) and is compared to NN & PSO and other method .The back propagation neural network has been used as classifier that its input are image pixels H,S and V features. In order to optimization the NN weight, the GA and PSO have been used. Dataset that has been used in this paper consists of 200 thousands skin and non-skin pixel that has been produced in HSV color-space. Result efficiency is 98.825% (accurate of correct identification) that is comparable to the other former methods. The advantage of this method is high rate and accuracy to identify skin in 2-dimentional images. Thus can use this method in real times. We compare accuracy and rate of the proposed method with the other known methods for show Verity of this work.**

Keywords- Hybrid NN& GA; Genetic Algorithm; PSO; HSV color-space; Back propagation

I. INTRODUCTION

Human skin is one of widespread theme in human image processing that present in many applications such as face detection[1] and the detection process of images with naked or scantily dressed people[2], commercial application, for example the driver eye tracker developed by forduk [3]. In images and videos, skin color is an indication of the existence of humans in such media. Therefore, in the last two decades extensive research have focused on skin detection in images. Skin detection means detecting image pixels and regions that contain skin-tone color. Most the research in this area has focused on detecting skin pixels and regions based on their color. Very few approaches attempt to also use texture information to classify skin pixels. Skin color as a cue to detect a face has several advantages: First, skin detection techniques can be both simple and accurate and second, the color dos not vary significantly with orientation or view angles, under white light conditions.

However, color is not a physical phenomenon. It is a perceptual phenomenon that is related to the spectral characteristics of electro-magnetic radiation in the visible wavelengths striking the retina [4]. One of skin detection step is choosing a suitable color space. In other work has been used different color-space such as RGB that is used by Rehg and Jones [5], HSI, HSV/HSB is used in [6], YUV, YIQ and etc. in this work is used HSV color-space. Next step is Choosing a classifier and learning .the classifiers are used in different work are Bayesian model, Gaussian model [7] and NN model. This work propose the hybrid NN and GA as classifier and is compared its result with other work, that detect better result than they.

The paper is organized as follows: Section 2 presents skin detection algorithm in this work. Section 3 explains the skin feature detection. Section 4 introduces the neural network (NN). Section 5 introduces the optimization algorithm (GA and PSO). Section 6 presents results and discussions. The final section gives conclusions.

II. SKIN DETECTION ALGORITHM

The purpose Skin detection algorithms can be classified into two groups: pixel-based [8] and context-based [9]. Since context-based methods are built on top of pixel-based ones, an improvement on a pixel-based methodology supposes a general advancement in the resolution of skin detection. Pixel-based algorithms classify each pixel individually without taking the other pixels of the image into consideration. These methodologies realize the skin detection either by bounding the skin distribution or by using statistical models on a given color space.

In this work is used pixel- based algorithm. Thus algorithm step are follows generally:

1. Collecting a database of 200 thousands skin and non-skin pixel
2. Choosing a suitable color-space (HSV in this work the advantages of these color spaces in skin detection is that they allow users to intuitively specify the boundary of the skin color class in terms of the hue and saturation). And converting the pixels into the HSV color- space.
3. Using neural network as classifier and Learning the weighs of neural network.
4. Optimization neural network weights using GA and PSO algorithm.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

5. testing given image (a. converting the image pixels into the HSV color space, b. classifying each pixel using the skin classifier to either a skin or non-skin)..

III. SKIN FEATURES DETECTION

Before Perceptual color spaces, such as HSI, HSV/HSB, and HSL (HLS), have also been popular in skin detection. These color spaces separates three components: the hue (H), the saturation (S) and the brightness (I, V or L). Essentially, HSV-type color spaces are deformations of the RGB color cube and they can be mapped from the RGB space via a nonlinear transformation as follow [10]:

$$H = \arccos \frac{{}^{1}/_{2}\left((R-G)-(R-B)\right)}{\sqrt{\left((R-G)^2-(R-B)(G-B)\right)}} \quad (1)$$

$$S = 1 - 3\frac{\min(R,G,B)}{R+G+B} \quad (2)$$

$$V = \frac{1}{3}(R+G+B) \quad (3)$$

One of the advantages of these color spaces in skin detection is that they allow users to intuitively specify the boundary of the skin color class in terms of the hue and saturation. As I, V or L give the brightness information, they are often dropped to reduce illumination dependency of skin color.

Considering low HSV color-space sensitivity versus white light intensity, brightness and surface orientation than light source in RGB to HSV converting, the HSV color-space is used for acquest skin features, in this paper. Thus HSV color space is proper to colored regions such as skin. First, RGB skin and non-skin pixel from dataset convert to the HSV color-space. After converting for each pixel obtain a three-dimension feature vector (H, S, V) as input for neural network.

IV. NEURAL NETWORK

Neural networks are non-linear classifiers and have been used in many pattern recognition problems like optical character recognition and object recognition. There is many image based face detection using neural networks [11] the most successful system was introduced by Rowley et al [12] as using skin color segmentation to test an image and classify each DCT based feature vector for the presence of either a face or non face.

The neural network used in this paper is back propagation neural network. Back propagation is a descent gradient search algorithm, which tries to minimize the total error square between actual output and target output of neural networks. This error is used to guide BP's search in the weight and bias space. There have been some successful applications of BP algorithms and use in artificial intelligence widely. However, there are drawbacks with the BP algorithms due to its descent nature. Studies show back propagation training algorithm is very sensitive to initializing conditions and often get trapped in local minimum of the function. To overcome those drawbacks, global search procedures like PSO and GA algorithms can be applied into the training process effectively. In this paper is applied the GA algorithm in order to optimization neural network weight.

There are two issues that must be addressed in design of a BP networks-based skin detector, the choice of the skin features (that has been described in previous section) and the structure of the neural networks. The structure defines how many layers the network will have, the size of each layer, the number of inputs of the network and the value of the output for skin and non-skin pixels. Then the network is trained using samples of skin and non-skin pixels. Considering to both of training time and ability of classifying the structure of the neural network is used in this work is adopted as figer.1.

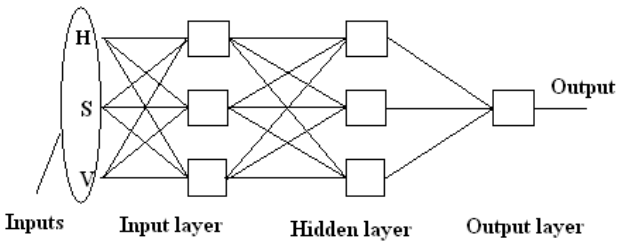


Figure 1. the neural networks structure

It has three layers, tree neuron in the input layer that its inputs are H, S and V feature for each skin or non-skin pixel from dataset, single neuron in output layer which detect the skin or non-skin pixels and tree neuron in hidden layer which is obtained by the experimental formula [13]:

$$n_i = \sqrt{n+m} + \alpha \quad (4)$$

Where n and m are the number of input and output neuron respectively. α is a constant between 1 and 10. Each neuron contains the weighted sum of its inputs filtered by a sigmoid (al) (s- shaped) transfer function:

$$f(x) = \frac{1}{1+e^{\sigma x}} \quad (5)$$

The parameter σ plays a very important role in the convergence of the neural networks: the larger σ is, the neural networks will converge more quickly, but also easy get unstable. On the other hand, if σ is too small, the convergence of the neural networks will be time consuming though. May get good result.

V. OPTIMIZATION ALGORITHM

A. *Genetic Algorithm*

GAs are search procedures which have shown to perform well considering large search spaces. We have used GA due to optimization weights and biases of neural network. The GA is described as follow:

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A chromosome in a computer algorithm is an array of genes. In this work each chromosome contains the array of 21weights and 7bias, that has an associated cost function assigned to the relative merit.

[Chromosome= (w1, w2,……w21,b1,b2,….b7)]

The algorithm begins with 50 initial population which chromosomes are generated randomly .min and max of each chromosome is obtained considering result weights and biases from NN, then cost function is evaluated for each chromosome. The cost function computes error for each chromosome using NN for training data. Error that is the same fitness is computed as 6 simple equation:

$$\text{Fitness} = \sum (Y_m \neq \overline{Y_m}) \quad (6)$$

Where Y_m is the target output for input data apply to NN and $\overline{Y_m}$ is the result output considering weights and biases accordance with the current chromosome. The population which is able to reproduce best fitness is known as parents. Then the GA goes into the production phase where the parents are chosen base on the least cost (best fitness is least cost because of we want the error be minimum). The selected parents reproduce using the genetic algorithm operator called crossover. In crossover random points are selected. When the new generation is complete, the process of crossover is stopped. Mutation has a secondary role in the simple GA operation. Mutation is needed because, even though reproduction and crossover effectively search and recombine extant notions, occasionally they may become overzealous and lose some potentially useful genetic material. After mutation has taken place, the fitness is evaluated. Then the old generation is replaced completely or partially. This process is repeated. After the algorithm reaches to minimum error or the iteration completed, it stops. The final chromosome is optimization weights and biases that are applied to neural network.

B. PSO Algorithms

In PSO algorithm, any solution that is called a particle is equivalent to a bird in the birds swarm motion pattern [14]. Any particle has a fitness which is computed by cost function. Whatever, any particle in searching area be close to objective-food (in birds model), it has the higher fitness. Also any particle has a velocity that lead to the particle motion. Particles follow the optimum particle and continue to the motion in problem space in each iteration.

The PSO Launches as: a Group of particles are generated accidentally (is considered 50 in this work), and by updating the generations, try to reach an optimum solution. In any step each particle using 2 best values are updated. The first case is the best condition that a particle has reached .The said position, is called "pbest" and is saved. Another best value that is used by algorithm is the best situation that has been acquired by the population so far. It is presented by "gbest".

$$V[] = v[] + c_1 * \text{rand}() * (\text{pbest}[] - \text{position}[]) + c_2 * \text{rand}() * (\text{gbest}[] - \text{position}[]) \quad (7)$$

$$\text{Position}[] = \text{position}[] + v[] \quad (8)$$

Where v [] is the particle velocity and position [] is the current particle position. They are arrays that their length is equal to problem dimensions. Rand () is a random number between 0 and 1. c_1 and c_2 are learning factors. In this article $c_1=c_2=0.5$.

The first step of applying PSO to training a neural network is to encode the solutions. In this article, any solution contains 28 parameters representing 21 weights and 7 biases for the neural networks:

[Chromosome= (w1,w2,……w21,b1,b2,….b7)]

The population value is considered 50 too. For each solution, the training set enter to the neural network and calculate the total system errors as 6 equation(cost function).and the algorithm performs as is described above. Final the best solution as optimum weights and biases enter to the neural network and is computed the correct rate for test data.

VI. RESULTS AND DISCUSSION

Proposed method is performed using MATLAB simulator. 200 thousand skin and non-skin pixels from 530 RGB image which have been collected from real and reliable training dataset [15] for learning the algorithm. The elements such as age, race, background, gender, light and brightness condition is considered in selecting image. For using trained network, in order to identify the skin pixel, first each RGB pixel convert to the HSV color space and Then H, S and V features apply to the trained network as the input. Afterwards, according to the output, the network classifies the pixel as skin or non-skin. The skin regions specify with white color and the non-skin regions with the black color. The criterion which we consider in this work is the correct rate. It is compute as follow:

Correct rate= ((length (target test) – error)/ length (target test))*100

the result of neural network performance at each time is different due to randomly initial weight .Thus we perform the NN three time, and its results associated whit GA and PSO are given in figure 2, 3 Figure 2 is obtained with 59.175%, 59.23 and 83.982% correct rate for NN, NN&PSO and NN &GA respectively and figure 3 with 70.6075%, 69.93% and 84.5%. the result show the NN& GA has the best result because of the GA spot the initial population base on min and max of the result of NN weight. But the PSO choose random the initial population completely. However, by the more performance, the better result with higher correct rate is obtained. We reach to 98.825% correct rate using this hybrid algorithm.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

To compare the proposed method with other techniques, Gaussian and Bayesian methods have modeled, and result of binary images, were presented in fig. 4 associated with result of proposed method. The fist column is original image, second column Gaussian method, third column Bayesian method, fourth column NN method and fifth column presents the proposed method (NN&GA). As it can be seen from the figure, the Gaussian and Bayesian methods ,has specified some points of background image as the skin wrongly and also NN method considered some cloths them as skin while the proposed method correctly presented skin regions. The

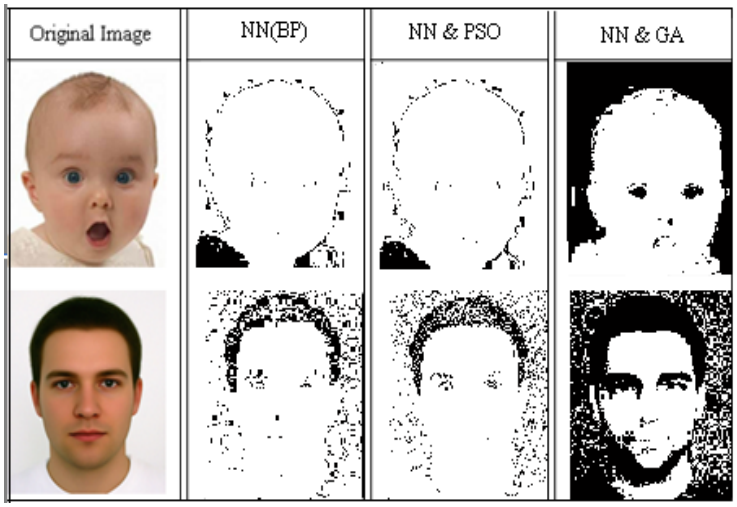


Figure 2. the result of simulation for NN, NN&GA and NN & PSO with 59.175%, 83.982% and 59.23%% correct rate respectively.

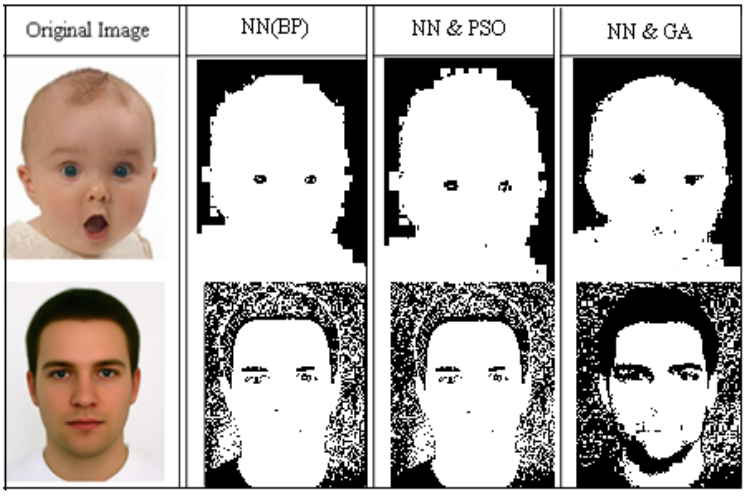


Figure 3. the result of simulation for NN, NN&GA and NN & PSO with 70.6075%, 84.5% and 69.93% correct rate respectively.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

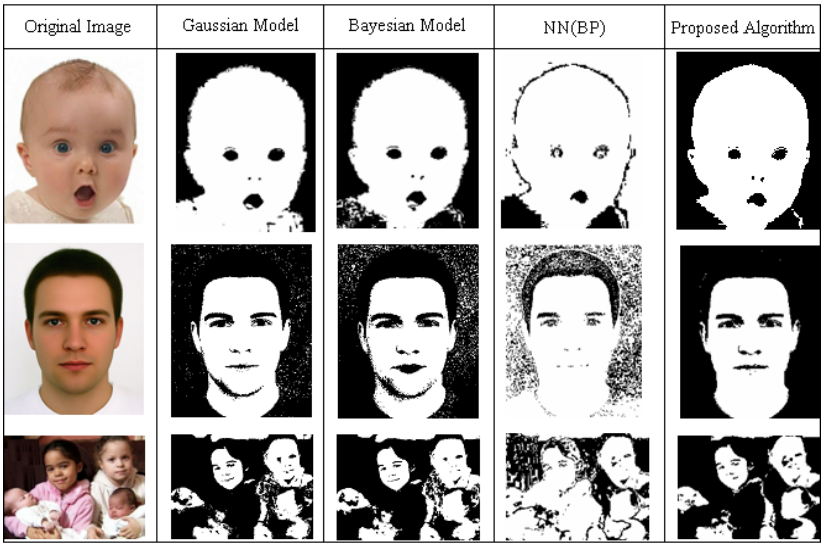


Figure 4. comparison of the proposed method against Gaussian, Bayesian and neural network with 98.825% correct rate.

VII. CONCLUSIONS

Skin detection, is an important preprocess in any analytical image regions. Accuracy is vital in post-processing. In this article, NN & GA hybrid method has presented for human skin detection. The experiment presented constant accuracy more than 98/825% on the human skin. HSV color space has been selected in this article, because it has lower sensitivity versus environmental condition and lightness. The various skin detection algorithms that have been presented so far, that they have advantages and disadvantages. One of the most important factors is time order of these techniques. As an example parzen method, is not analogous with methods like Gaussian and Bayesian. Despite having the very down time order, the proposed method, present reliable results compare to previous methods.

REFERENCES

[1] I C. A. Bouman. Cluster: AN Unsupervised Algorithm for Modelling Gaussian Mixture. School Of Electrical Engineering, Purdue University, http://dynamo.ecn/ bouman/software/cluster,September 2000.

[2] Fleck, M.M., Forsyth, D.A., Bregler, C.: Finding naked people. In: Proceedings of the European Conference on Computer Vision (ECCV). (1996) 593–602

[3] J D.Tock and I.Craw. Tracking and measuring drivers' eyes. Image and Vision Computing, 14:541-548,1996.

[4] J. Yang and A. Waibel. " A real-time fase traker". In Processing of the third workshop on Application of Computer Vision, pp 142-147, 1996

[5] Jones, M.J., Rehg, J.M.: Statistical color models with application to skin detection. International Journal of Computer Vision (IJCV) 46(1) (2002) 81–96.

[6] Albiol, A., Torres, L., Delp, E.: Optimum color spaces for skin detection. In: Proceedings of the International Conference on Image Processing (ICIP). (2001) I: 122–124

[7] Yang, M., Ahuja, N.: Gaussian mixture model for human skin color and its application in image and video databases. In: In Proc. of the SPIE: Conference on Storage and Retrieval for Image and Video Databases (SPIE 99). Volume 3656. (1999) 458466.

[8] J.Yang,W. Lu, A.Waibel, Detecting human faces in color images, in: IEEE Internat. Conf. on Image Processing, vol. 1, 1998, pp. 127–130.

[9] T.-Y. Cow, K.-M. Lam, Mean-shift based mixture model for face detection on color image, in: IEEE Internat. Conf. on Image Processing, vol. 1, 2004, pp. 601–604.

[10] Vladimir Vezhnevets, Vassili Sazonov, Alla Andreeva" A Survey on Pixel-Based Skin Color Detection Techniques". †www: http://graphics.cmc.msu.ru

[11] Lamiaa Mostafa and Sherif Abdelazeem" Face Detection Based on Skin Color Using Neural Networks" GVIP 05 Conference, 19-21 December 2005, CICC, Cairo, Egypt.

[12] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," in Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, 1998, pp. 23-38.

[13] W. Kelly, A. Donnellan, D. Molloy, 'Screening for Objectionable Images: A Review of Skin Detection Techniques', © 2008 IEEE

[14] Kennedy, J. and Eberhart, R. C., "Particle Swarm Optimization", Proceedings of IEEE International Conference on Neural Networks, Piscataway, NJ, pp. 1942-1948, 1995

[15] http://Ibmedia.ece.ucsb.edu/resources/dataset

HAND GEOMETRY IDENTIFICATION BASED ON MULTIPLE-CLASS ASSOCIATION RULES

A.S.Abohamama[1], O.Nomir[2], and M.Z. Rashad[3]
[1, 2, 3] Department of Computer Sciences, Mansoura University, Mansoura, Egypt
Emails: [1] Aabohamama@yahoo.com, [2] o.nomir@umiami.edu, [3] Magdi_12003@yahoo.com

***Abstract*- Hand geometry has long been widely used for biometric verification and identification because of its user acceptance, its good verification, and its identification performance. In this paper, a biometric system is presented for controlled access using hand geometry. It presents a new approach based on multiple-class association rules (CMAR) for classification. The system automatically extracts a minimal set of features which uniquely identify each single hand. CMAR is used to build the identification system's classifier. During identification, the hands that have features closer to a query hand are found and presented to the user. Experimental results using a database consists of 400 hand images from 40 individuals are encouraging. The proposed system is robust, and a good identification result has been achieved.**

Keywords: Biometric systems; Hand Geometry; CMAR; Classification.

i. INTRODUCTION

A biometric system is able to identify an individual based on his / her physiological traits such as fingerprint, iris, hand and face. It also can identify an individual based on behavioral traits such as gait, voice and handwriting [1].

Biometric techniques differ according to security level, user acceptance, cost, performance, etc. One of the physiological characteristics for individual's recognition is hand geometry.

Each biometric technique has its own advantages and disadvantages. While some of them provide more security, i.e. lower False Acceptance Rate (FAR) and False Rejection Rate (FRR), other techniques are cheaper or better accepted by the final users [2].

Hand geometry identification is based on the fact that the hand for any individual is unique. In any individual's hand, the length, width, thickness, and curvatures for each finger as well as the relative location of these features distinguish human being from each other [3]. As often noted in the literature, hand shape biometrics is interesting to study due to the following reasons [4]:

1) Hand shape can be captured in a relatively user convenient, non-intrusive manner by using inexpensive sensors.
2) Extracting the hand shape information requires only low resolution images and the user templates can be efficiently stored (nine-byte templates are used by some commercial hand recognition systems).
3) This biometric modality is more acceptable to the public mainly because it lacks criminal connotation.
4) Additional biometric features such as palm prints and finger-prints can be easily integrated to an existing hand shape-based biometric system.

Environmental factors such as dry weather or individual anomalies such as dry skin do not appear to have any negative effects on the verification accuracy of hand geometry-based systems. The performance of these systems might be influenced if people wear big rings, have swollen fingers or no fingers. Although hand analysis is most acceptable, it was found that in some countries people do not like to place their palm where other people do. Sophisticated bone structure models of the authorized users may deceive the hand systems. Paralyzed people or people with Parkinson's disease will not be able to use this biometric method [3].

In the literature, there are some techniques using different features used for hand geometry's identification [1] [3] [5] [6] [7] [8].

In [1], they presented an approach to automatically recognize hand geometry pattern The input hand images were resized and converted to a vector before they are applied to the input of the General regression neural networks (GRNN) for hand geometry identification The system does not require any feature extraction stage before the identification.

In [3], they transformed the hand images to binary images, removed the image's noise, and extracted the hand boundary. The extracted features are the widths of the fingers and they are measured in three different heights (i.e. measured at three different locations) except the thump is measured in two heights, the lengths of all fingers, and two measurements of the palm size. The result is a vector of 21 elements is used to identify persons. Euclidian distance, Hamming distance, and Gaussian mixture model are used for classification.

In [5], they binarized the hand image and extracted two completely different sets of features from the images. The first set is geometric measurements consist of 10 direct features; they are the length of the fingers, three hand ratio measurements, area, and perimeter. The second set is the hand contour information. In order to reduce the length of the template vector, they used the Principal Component Analysis (PCA), wavelet transform, and cosine transform. The classification techniques used are multilayer perceptron

* Corresponding Author
Name : A.S.Abohamama
Mail: Aabohamama@yahoo.com
Tel: 020141641771

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

neural network (NNMLP) and nearby neighbor classifier (KNN).

In [6], they used the palm print and hand geometry features for identification. The extracted features are the hand's length, width, thickness, geometrical composition, shape and the geometry of fingers, and shapes of the palm etc. The extracted palm print features are composed of principle lines, wrinkles, minutiae, delta points, etc. These features are grouped into four different feature vectors. A K-NN classifier based on majority vote rule and distance weighted rule is employed to establish four classifiers. Dempster-shafer evidence theory is then used to combine these classifiers in case of identification.

In [7], they proposed a hierarchical identification method based on improved hand geometry and regional content features for low resolution hand images without region of interest's (ROI) cropping. At coarse levels, angle information is added as a complement to line-based hand geometry. At fine levels, relying on the assumption that gradient value of each pixel presents the gray-level changing rate. They developed a simple sequence labeling segmentation method, and chose conditional regions that are relatively steady in segmentation through region area constraint. Because distinctive lines and dense textures always have lower gray-levels than their surrounding areas, regions with lower average gray-levels are selected from conditional regions. Regional centroid coordinates are extracted as feature vectors. Finally, regional spatiality relationship matrix is built up to measure distances between feature vectors with various dimensions.

In [8], the palm prints and hand geometry images are extracted from a hand image in a single shot at the same time. To extract the hand geometry features, each image is binarized and aligned to preferred direction. The geometry features are the length, the width of fingers, the palm width, the palm length, the hand area, and the hand length. The ROI method is issued to extract the palm print images. The extracted palm print images are normalized to have prespecified mean and variance. Then significant line features are extracted from each of the normalized palm print images. Matching score level fused with max rule are used for classification.

The aim of our work is to develop a simple and effective recognition system for identifying individuals using their hands' features. The proposed identification process relies on extracting a minimal set of features which uniquely identify each single hand. The CMAR technique is used to build the classifier of our identification system. The block diagram of the proposed identification system is shown in Fig.1.

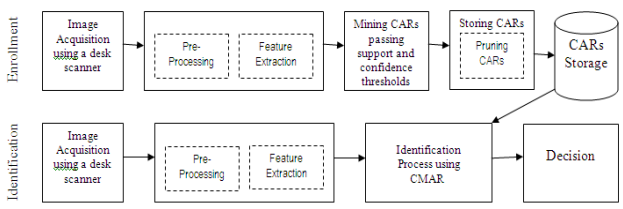


Figure 1. Block diagram of a Biometric Recognition System.

In our proposed system and during the enrollment, a set of samples are taken from the users, and some features are extracted from each sample. During the training step, the extracted features that represent the training data set are used in the generation of Class Association Rules (CARs) which will be pruned depending on specific criteria yielding our classifier. After the training step is completed, the classifier is stored in an efficient data structure. Given a user who wants to gain access, a new sample is taken from this user and the sample's features are extracted. The extracted feature vector is then used as an input to the previously stored classifier. Then, the obtained output is analyzed and the system decides if the sample belongs to a user previously enrolled in the system or not. Our identification procedure is described in the following sections. Our paper is organized as follows, Section two presents preliminaries about the proposed technique, Section three presents feature extraction, Section four presents how Multiple- Class Association Rules are used in hand geometry classification, Section five presents experimental result, and finally Section six concludes the paper.

ii. PRELIMINARIES

A. Hand geometry and Image Acquisition

Hand geometry has long been used for biometric verification and identification because of its acquisition convenience and good verification and identification performance. From anatomical point of view, human hand can be characterized by its length, width, thickness, geometrical composition, shapes of the palm, and shape and geometry of the fingers. Earlier efforts in human recognition used combinations of these features with varying degrees of success. The hand images can be taken in two ways in which hand position is either controlled with pegs or not. Traditionally, pegs are almost always used to fix the placement of the hand, and the length, width and thickness of the hand are then taken as features [9].

Pegs will almost definitely deform the shape of the hand. Even though the pegs are fixed, the fingers may be placed differently at different instants, and this causes variability in the hand placement. These problems will degrade the performance of hand geometry verification because they adversely affect the features [9].

Without the needs for pegs, the system has simple acquisition interface. Users can place their hands in arbitrary fashion and can have various extending angles between the five fingers. The Main points are then extracted from the segmented image and used to compute the required features.

In our system, we used a database consisting of 10 different acquisitions of 40 people. They have been taken from the users' right hand. Most of the users are within a selective age range from 23 to 30 years old. The percent of males and females are not equal. The images have been acquired with a typical desk-scanner using eight bits per pixel (256 gray levels), a resolution of 150 dpi. (Available

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

in: <http://www.gpds. ulpgc.es/download>) [1][10].Some images are shown in Fig. 2**.**

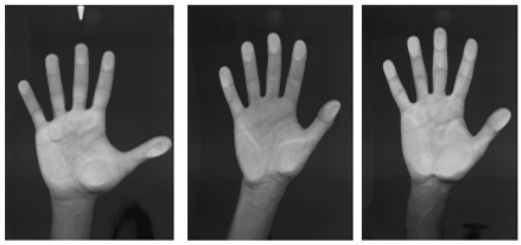


Figure 2. Templates captured by a desk scanner.

B. Classification Based on Multiple-Class Association Rules (CMAR)

Given a set of cases with class labels as a training set, a classifier is built to predict future data objects for which the class label is unknown [11]. In other words, the purpose of the classification step is to identify a new data in virtue of current knowledge as more as possible [12].

In our work we use a special type of classification called associative classification. Associative classification, one of the most important tasks in data mining and knowledge discovery, builds a classification system based on associative classification rules [13].

Associative classification techniques employ association rule discovery methods to find the rules [14]. This approach was introduced in 1997 by Ali, Manganaris, and srikant. It produced rules for describing relationships between attribute values and the class' attribute. This approach was not for prediction, which was the ultimate goal for classification in 1998, associative classification has been employed to build classifiers [14].

CBA, classification based on associations (Liu, Hsu, & Ma, 1998), is an algorithm for building complete classification models using association rules. In CBA, all class association rules are extracted from the available training dataset (i.e., all the association rules containing the class attribute in their consequent). The most suitable rules are selected to build an ''associative classification model", which is completed with a default class [13].

Extensive performance studies show that association based classification may have better accuracy in general. However, this approach may also suffer some weakness because of some reasons. **First,** *it is not easy to identify the most effective rule at classifying a new case* so many methods select a single rule with a maximal user-defined measure, such as confidence. Such a selection may not always be the right choice in many cases. **Second,** *a training data set often generates a huge set of rules. It is challenging to store, retrieve, prune, and sort a large number of rules efficiently for classification [11]*.

CMAR, Classification based on Multiple Association Rules, developed basically to overcome the previous problems related to association based classification. In CMAR, instead of relying on a single rule to classify data, CMAR considers sets of related rules, taking into account that the most confident rule might not always be the best choice for classifying data. Given a data object, CMAR retrieves all the rules matching that object and assigns a class label to it according to a weighted $\chi 2$ measure, which indicates the ''combined effect" of the rules. Also, CMAR adopts a variant of the FP-growth algorithm to obtain and efficiently store rules for classification in a tree structure [13].

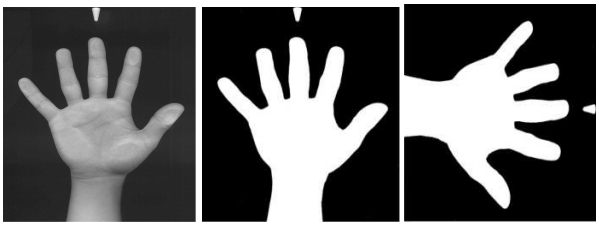
CMAR consists of two phases: rule generation and classification. In the first phase, rule generation, CMAR computes the complete set of rules in the form of R: P ➔ C, where P is a pattern in the training data set and C is a class label such that Sup(R) and Conf(R) pass the given support and confidence thresholds, respectively. Furthermore, CMAR prunes some rules and only selects a subset of high quality rules for classification [11].

In the second phase, classification, for a given data object *obj*, CMAR extracts a subset of rules matching the object and predicts the class label of the object by analyzing this subset of rules [11].

iii. FEATURE EXTRACTION

A. Image Preprocessing

After the image is captured, it is preprocessed to obtain only the area information of the hand. The first step in preprocessing is to transform the hand image to binary image. Since there is clear distinction in intensity between the hand and the background, the image can be easily converted to a binary image by thresholding. The result of the binarization step for the image in Fig. 3a is shown in Fig. 3b. After the completion of binarization process the bianarized image is rotated counterclockwise by 270 degrees. The rotated image is shown in Fig. 3c.



(a) (b) (c)

Figure 3. Image binarization and rotation (a) Original Image (b) The binary image (c) The rotated binary image

The next step in the preprocessing is obtaining the boundary of the binary hand image. Fig. 4 shows the result of extracting the hand's boundary for the binary image in Fig. 3c.

Figure 4. The boundary captured for the binary hand image in Fig. 3c

.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. Extracting the Features

We implement an algorithm for feature extraction. The algorithm is based on counting pixel distances in specific areas of the hand. The first step in extracting features is to measure the main points (finger tips and valleys between fingers), there are shown in Fig. 5.

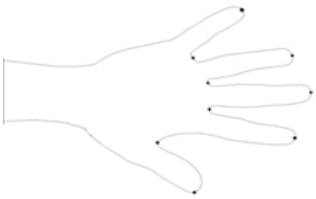


Figure 5. Capturing the main points

From these main extracted points, we locate all other points required to calculate our features vector. The algorithm looks for white pixels between two located points and computes a distance using geometrical principles. The calculated features vector consists of 16 different values, as follows:

- *Widths:* each of the fingers is measured in 2 different heights. Thump finger is measured in one height.
- *Lengths:* the lengths of all fingers and thumb are obtained.
- *Palm:* one measurements of palm size.
- Distance from the middle finger's tip to the middle of the palm.

The extracted features for the located main points in Fig. 5 are shown in Fig 6. Then each length is divided by a width, in other words the length of each finger is divided by the different widths of that finger and the distance from the middle finger's tip to the middle of the palm is divided by the palm width, to handle the aspect ratio problem. The result is a vector of only 12 elements.

Figure 6. The extracted features for the located main points in Fig. 5.

iv. CMAR IN HAND GEOMETRY CLASSIFICATION

We now will make an overview of how CMAR algorithm works. For more detail CMAR algorithm discussed at [11] and [15].

CMAR is a Classification Association Rule Mining (CARM) algorithm developed by Wenmin Li, Jiawei Han and Jian Pei (Li et al. 2001). CMAR operates using a two stage approach to generate a classifier [15]:

1. Generating the complete set of CARs according to a user supplied:
   a. Support threshold to determine frequent item sets, and
   b. Confidence threshold to confirm CRs.
2. Prune this set to produce a classifier.

CMAR algorithm uses FP-growth method to generate a set of CARs which are then stored in an efficient data structure called CR-tree. CARs are inserted in the CR-tree [15] if:

1. CAR has Chi-Squared value above a user specified critical threshold.
2. The CR tree does not contain a rule that have a higher rank.

Given two CARs, R1 and R2, R1 is said having higher rank than R2 [11] if:

1- If confidence(R1) > confidence(R2).
2- If confidence(R1) == confidence(R2) && support(R1) > support(R2).
3- If confidence(R1) == confidence(R2) && support(R1) == support(R2) but R1 has fewer attribute values in its left hand side than R2 does.

After the production of the CR-tree the set of CARS are pruned based on the cover principle meaning that each record is covered by N CAR. We used LUCS-KDD implementation of CMAR in which the threshold for Chi-Squared test is `3.8415` and N = 3.

To test the resulting classifier given a record *r* in the test set collect all rules that satisfy *r*, and
1. If consequents of all rules are all identical classify record according to the consequents
2. Else group rules according to classifier and determine the combined effect of the rules in each group. The classifier associated with the "strongest group" is then selected.

The strength of a group is calculated using a Weighted Chi Squared (WCS) measure [11].

v. EXPERIMENTAL RESULTS

The aim of our work is to develop a simple and effective recognition system to identify individuals using their hand's geometry. We proposed a new technique using the CMAR to build the classifier to classify individuals using their hand features.

Our database contains a set of different hand images. This database has been built off-line using a desk scanner [10]. It contains 400 samples taken from 40 different users.

The database is then pre-processed in order to prepare the images for the feature extraction phase. This process is composed by three main stages: binarization, contour and

.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

main points extraction (finger tips and valleys between fingers), and then we extracted a minimal set of features, 12 values uniquely identify each person's hand. These extracted features are later used in the recognition process. These features included the length of fingers, the width of fingers, and the width of the palm. These features are archived along with the hand images in the database.

We utilized LUCS-KDD implementation of CMAR to build the classifier which consists of CARs stored in an efficient data structure referred to as CR-tree.

Given a query hand, our system applies the preprocessing stage to this input query hand. Then, the feature vector for this query hand is extracted. This extracted feature vector is presented as an input to the CMAR classifier which collects all rules that satisfy the feature vector, and if consequents of all rules are all identical then the feature vector is classified according to the consequents else rules are grouped according to the class (consequent of the rule) and the combined effect of the rules in each group is determined, the class associated with the "strongest group" is then selected.

Original LUCS-KDD implementation of CMAR takes all dataset as input (training data and test date) and uses a 50:50 training/test set split. We modified LUCS-KDD implementation of CMAR to take 8 samples for training and 2 samples for test. The support threshold and confidence threshold are 1 and 50 respectively.

Our system performance is measured using identification rate, and the results are shown at Table1.

TABLE I. Identification rate values for some experiments using different number of persons.

Number of Persons	Identification Rate
10 person	96.70%
20 person	95.32%
30 person	94.67%
40 person	94.01%

We compared our identification results with the identification results in [1]. In [1], during the enrollment stage, seven images for each person were used for training and three images different from training images were used for testing. For the intruders, two images for each person were used for validation. For hand geometry identification, the application is carried out for 20 authorized users and considerable identification rate is obtained. Their proposed model achieved 93.3% in testing (test stage is realized for authorized users). Comparing to our identification results, our identification results is considered better Also, our dataset is larger than their dataset, i.e. the number of enrolled subject.

vi. CONCLUSION

In this paper, we presented a biometric system using hand geometry. A new approach using CMAR is presented to build the identification system's classifier. Our system automatically extracts a minimal set of features which uniquely identify each single person's hand. During archiving, the features are extracted and stored at the database along with the images. During identification, the hands that have features closer to a query hand are found and presented to the user. Experimental results on a database consists of 400 hand images from 40 individuals are encouraging. We have shown experimental results for images of different qualities. We use the identification rate to measure the system performance. The experimental results prove that the proposed system is robust, and a good identification result has been achieved**.** We compared the performance of our proposed identification system with the system introduced in [1], our proposed system outperforms that identification systems in terms of identification rate.

Acknowledgment

We are so thankful to Miguel A. Ferrer for providing us with the used dataset.

REFERENCES

[1]Övünç Polat , Tülay Yıldırım , "Hand geometry identification without feature extraction by general regression neural network " , An International Journal in Expert Systems with Applications, vol 34, No 2, February 2008, pp. 845-849.

[2]R. Sanchez-Reillo, "Hand Geometry Pattern Recognition through Gaussian Mixture Modeling", *15th International Conference on Pattern Recognition (ICPR'00)*, vol 2, September 2000, pp. 29-39

[3]P.VARCHOL, D. LEVICKY, "Using of Hand Geometry in Biometric Security Systems", *Radioengineering*, vol 16, No. 4, December 2007, pp. 82-87.

[4] Nicolae Duta , A survey of biometric technology based on hand shape, *Elsevier Ltd , Pattern Recognition*, vol 42, No.11 , November 2009, pp. 2797-2806.

[5] S. Gonzalez, C.M. Travieso, J.B. Alonso, M.A. Ferrer, "Automatic biometric identification system by hand geometry", *IEEE 37th Annual International Carnahan Conference in Security Technology*, October 2003, pp. 281-284.

[6]M. Arif, T. Brouard, N. Vincent, KRL, Rawalpindi, "Personal Identification and Verification by Hand Recognition", *IEEE International Conference in Engineering of Intelligent Systems*, September 2006, pp. 1-6.

[7] Jie Wu , Zhengding Qiu , Dongmei Sun, "A hierarchical identification method based on improved hand geometry and regional content feature for low-resolution hand images", *Elsevier North-Holland, Inc. Signal Processing* , vol 88, No. 6 , June 2008, pp. 1447-1460

[8] Ajay Kumara, David C.M. Wongb, Helen C. Shenb, and Anil K. Jainc, , "Personal authentication using hand images", Elsevier Science Inc. New York, NY, USA, *Pattern Recognition Letters*, vol 27, No.13, October 2006, pp.1478-1486.

[9] Alexandra L.N.wong, Pengcheng shi ,"Peg-Free Hand Geometry Recognition Using Hierarchical Geometry and Shape Matching", *IAPR Workshop on Machine Vision Applications*, Nara- ken New Public Hall, Nara, Japan, December 2002, pp. 281 - 284.

[10]Miguel A. Ferrer, Aythami Morales, Carlos M. Travieso, Jesús B. Alonso, "Low Cost Multimodal Biometric Identification System based on Hand Geometry, Palm and Finger Textures", in 41st Annual IEEE

International Carnahan Conference on Security Technology, ISBN: 1-4244-1129-7, Ottawa, Canada, October 2007, pp. 52-58.

[11] Wenmin Li, Jiawei Han, and Jian Pei, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules", First IEEE International Conference on Data Mining (ICDM'01), November 29-December 02, 0-7695-1119-8.

[12]Huawen Liu, Jigui Sun, and Huijie Zhang, "Post-processing of associative classification rules using closed sets", Expert Systems with Applications international journal, Volume 36 Issue 3, April 2009.

[13] Yen-Liang Chen , Lucas Tzu-Hsuan Hung, "Using decision trees to summarize associative classification rules", Expert Systems with Applications international journal, Volume 36 Issue 2, March 2009.

[14] Fadi Thabtah, Peter Cowling, "Mining the data from a hyperheuristic approach using associative classification", Expert Systems with Applications international journal, Volume 34 Issue 2, February 2008.

[15]http://www.csc.liv.ac.uk/~frans/KDD/Software/CMAR/cmar.htm, Created and maintained by Frans Coenen, Last updated 5 May 2004

.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Survey on Web Usage Mining: Pattern Discovery and Applications

Ms.C. Thangamani, Research Scholar
Mother Teresa Women's University
Kodaikanal
thangamanic@yahoo.com

Dr. P. Thangaraj, Prof. & Head
Department of computer Science & Engineering
Bannari Amman Institute of Technology, Sathy

Abstract--- The past decade is described by an unexpected development of the Web both in the quantity of Web sites and in the quantity of the accessing users. This enlargement generated huge quantities of data related to the user interaction with the Web sites, recorded in Web log files. In addition, the Web sites holders uttered the requirement to recognize their visitors in an effective way so as to provide them web sites with satisfaction. The Web Usage Mining (WUM) is developed in recent years in order to discover knowledge from databases. WUM consists of three phases: the preprocessing of raw data, the discovery of schemas and the analysis of results. A WUM technique gathers usage behavior from the Web usage data. Large amount of web usage data makes difficulty in analyzing those data. When applied to large quantity of data, the existing techniques of data mining, usually, results in unsatisfactory outcome by means of behaviors of the Web sites' users. This paper focuses on analyzing the various web usage mining techniques. This analysis will help the researchers to develop a better technique for web usage mining.

Keywords--- Web Usage Mining, World Wide Web, Pattern Discovery, Data Cleaning

1. INTRODUCTION

Web Usage Mining is a component of Web Mining, which of course is a part of Data Mining technique. Since Data Mining includes the idea of mining significant and precious data from huge quantity of data, Web Usage mining includes extraction of the access patterns of the users in the web site. This gathered data can then be utilized in a various ways like improvement of the application, checking of fraudulent elements etc.

Web Usage Mining [16, 17] is usually referred as an element of the Business Intelligence in a business instead of technical characteristic. It is utilized for predicting business plans by means of the well-organized usage of Web Applications. It is also essential for the Customer Relationship Management (CRM) as it can guarantee customer fulfillment till the interaction among the customer and the organization is disturbed.

The main difficulty with Web Mining in general and Web Usage Mining in particular is the kind of data involved in processing. With the increase of Internet usage in this present world, the Web sites increased largely and a bundle of transactions and usages are happening by the seconds. Away from the quantity of the data, the data is not entirely ordered. It is organized in semi-structured manner so that it requires more preprocessing and parsing before the gathering of the necessary data from the entire data.

Web Data

In Web Usage Mining [18], data can be gathered from server logs, browser logs, proxy logs, or obtained from an organization's database. These data collections vary by means of the place of the data source, the types of data available, the regional culture from where the data was gathered, and techniques of implementation.

There are various kinds of data that can be utilized in Web Mining.

i. *Content*

ii. *Structure*

iii. *Usage*

Data Sources

The data sources utilized in Web Usage Mining may include web data repositories such as:

Web Server Logs – These are logs which contain the pattern of page requests. The World Wide Web Consortium preserves a regular arrangement for web server log files, but other informal designs are also subsist. Latest entries are characteristically affixed to the ending of the file.

Information regarding the request which includes client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are normally included. This information can be gathered into a single file, or split into separate logs like access log, error log, or referrer log. On the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

other hand, server logs usually do not gather user-specific data. These files are typically not available to regular Internet users. It can be accessible only by webmaster or other administrative individuals. A numerical examination of the server log may be utilized to gather traffic behavior by time of day, day of week, referrer, or user agent.

Proxy Server Logs - A Web proxy is a caching method which happens among client browsers and Web servers. It assists to decrease the load time of Web pages and also the network traffic at both the ends (server and client). A proxy server log includes the HTTP requests which are performed by various clients. This may serve as a data source to discover the usage pattern of a group of unspecified users, sharing same proxy server.

Browser Logs – Different browsers such as Mozilla, Internet Explorer etc. can be altered or different JavaScript and Java applets can be utilized to gather client side information. This execution of client-side data gathering needs user assistance, either in executing the working of JavaScript and Java applets, or to willingly utilize the altered browser. Client-side gathering scores over server-side gatherings as it decreases both the bot and session detection difficulties.

Web log mining usually involves the following phases:

- Preprocessing
- Pattern Discovery
- Pattern Analysis

This paper focuses on analysis about the various existing techniques with the phases described above.

2. Related Works

Web usage mining and statistical examinations are two methods to estimate practice of Web site. With the help of Web usage mining techniques, graph mining envelops complex Web browsing patterns like parallel browsing. With the help of statistical examination techniques, examining page browsing time suggests valuable data about Web site, usage and its users. Heydari *et al.[1],* suggested a graph-based Web usage mining technique which merges Web usage mining and statistical examination taking into account of client side data. Conversely, it merges graph based Web usage mining and browsing time examination by considering client side data. It assists the web site owners to predict the user session accurately and enhance the website. It is determined to predict the Web usage patterns with more accuracy.

Web usage mining is a technique of data mining in order to mine the information of the Web server log file. It can determine the browsing behaviors of user and some type of correlations among the web pages. Web usage mining offers the assistance for the Web site design, suggesting personalization server and other business making decision, etc. Web mining utilizes the data mining called the artificial intelligence and the chart expertise and so on to the Web data and outlines the users visiting characteristics, and then obtains the users browsing patterns. Han *et al.,* [2] performed a study on Web Mining Algorithm based on Usage Mining and it also constructs the design attitude of the electronic business website application technique. This technique is uncomplicated, efficient and effortless to understand and appropriate to the Web usage mining requirement of building a low budget website.

Web usage mining takes advantage of data mining methods to extract valuable data from usage behavior of World Wide Web (WWW) users. The required characteristics is captured by Web servers and stored in Web usage data logs. The initial stage of Web usage mining is the pre processing stage. In the preprocessing stage, initially, irrelevant data is cleared from the logs. This preprocessing stage is an important process in Web usage mining. The outcome of data preprocessing is appropriate to the further processing like transaction identification, path examination, association rule mining, sequential pattern mining, etc. Inbarani *et al.,* [3] proposed rough set based feature selection for Web log Mining. Feature extraction is a preprocessing phase in web usage mining, and it is highly efficient in decreasing the high dimensions to low dimensions by means of removing the irrelevant data, escalating the learning accuracy and enhancing comprehensiveness.

Web usage mining has grown to be fashionable in different business fields associated with Web site improvement. In Web usage mining, frequently interested navigational behavior are gathered by means of Web page addresses from the Web server visit logs, and the patterns are used in various applications including recommendation. The semantic data of the Web page text is usually not integrated in Web usage mining. Salin *et al.,* [4] proposed a structure for semantic information for web usage mining based recommendation. The repeated browsing paths are gathered by means of ontology instances as a substitute of Web page addresses and the outcome is utilized for creating Web page suggestions to the user. Additionally, an evaluation mechanism is implemented in order to test the accomplishment of the prediction. Experimental outcome suggests that highly precise

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

prediction can be resulted by considering semantic data in the Web usage mining.

In Web Usage Mining, web session clustering involves a major role to categorize web users in accordance with the user browsing behavior and similarity measure. Web session clustering in accordance with swarm assists in various manners to handle the web resources efficiently link web personalization, layout alteration, website alteration and web server performance. Hussain *et al.,* [5] proposed a hierarchical cluster based preprocessing methodology for Web Usage Mining. This structural design will envelop the data preprocessing phase to organize the web log data and translate the uncompromising web log data into mathematical information. A session vector is generated, in order that suitable resemblance and swarm optimization could be utilized to group the web log information. The hierarchical cluster based technique will improve the conventional web session methods for more structured data about the user sessions.

Mining the information of the Web server log files, determine the session behavior of user and several types of correlations among the Web pages. Web usage mining offers the assistance for the Web site creation, given that personalization server and additional business building judgment. There are various session regarding navigations are stored in Web server log files, page attribute of which is Boolean quantity. Fang *et al.,* [6] suggested a double algorithm of Web Usage Mining based on sequence number for the purpose of improving the effectiveness of existing technique and decrease the executing time of database scan. This is highly suitable for gathering user browsing behaviors. This technique modifies the session pattern of user into binary, and then utilizes up and down search approach to double generate candidate frequent itemsets. This technique calculates support by sequence number dimension with the purpose of scanning session pattern of user, which varies from existing double search mining technique. The evaluation represents that the proposed system is faster and more accurate than existing algorithms.

Huge quantity of information are collected repeatedly by Web servers and stored in access log files. Examination of server access log can afford considerable and helpful data. Web Usage Mining is the technique of utilizing data mining process to the identification of usage patterns from Web data. It analyses the secondary data obtained from the behavior of the users during some phase of Web sessions. Web usage mining composes of three stages such as preprocessing, pattern discovery, and pattern examination. Etminani *et al.,* [7] proposed a web usage mining technique for discovery of the users' navigational patterns using Kohonen's Self Organizing Map (SOM). Author suggests the usage of SOM to pre-processed Web logs using the web log collected from http://www.um.ac.ir/ and gathers the frequent patterns.

The web usage mining [19] makes use of data mining approaches to find out interesting usage patterns from the available web data. Web personalization utilizes web usage mining approaches for the development of customization. Customization concerns about knowledge acquisition through the analysis of user's navigational activities. A user when goes online more likely to obtain the links which is appropriate for his necessities or usage in the website he browses. The subsequent business requirement in the online industry will be personalizing/customizing the web page satisfying for each individuals need. The personalization of the web page will engage clustering of several web pages having general usage pattern. As the size of the cluster goes on mounting because of the increase in users or development of interest of users it will become inevitable requirement for optimizing the clusters. Alphy Anna *et al.,* [8] develops a cluster optimizing methodology in accordance with ants nestmate recognition capability and is used for removing the data redundancies that possibly will take place after the clustering done by the web usage mining techniques. For purpose of clustering an ART1-neural network based technique is used. "AntNestmate approach for cluster optimization" is presented to personalize web page clusters of target users.

Internet has turn out to be an essential tool for everyone, Web usage mining [20] in the same way becomes a hotspot, which uses huge amounts of data in the Web server log and further significant data sets for mining analysis and achieves valuable knowledge model about usage of important Web site. Several researches have to be done with the positive association rules in Web usage mining, however negative association rules is more significant, as a result Yang Bin *et al.,* [9] have applied negative association rules to Web usage mining. Experimental results have revealed that the negative association rules have a significant role on access pattern to Web visitors to resolve the troubles in which positive association rules are referred to.

Web usage mining (WUM) is a kind of Web mining, which utilizes data mining techniques to obtain helpful information from navigation pattern of Web users. The data must be preprocessed to enhance the effectiveness and simplify the mining process. Therefore it is significant to define before applying data mining techniques to determine user access patterns from Web log. The major use of data preprocessing is to prune noisy and unrelated data, and to lessen data volume for the pattern discovery stage. Aye *et al.,* [10] chiefly concentrates on data preprocessing stage of the initial phase of

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Web usage mining with activities like field extraction and data cleaning techniques. Field extraction techniques carry out the process of separating fields from the single line of the log file. Data cleaning technique removes inconsistent or unwanted items in the analyzed data.

The Internet is one of the rapidly growing fields of intelligence collection. When the users browse the website, the users leave a lot of records of their actions. This enormous amount of data can be a valuable source of knowledge. Sophisticated mining processes are required for this knowledge to extract, recognize and to utilize effectively. Web Usage Mining (WUM) systems are purposely designed to perform this task by examining the data representing usage data about a specific Web site. WUM can represent user behavior and, consequently, to predict their future navigation. Online prediction is the one of the major Web Usage Mining applications. On the other hand, the accuracy of the prediction and classification in the existing structural design of predicting users' future needs cannot still satisfy users particularly in large Web sites. In order to offer online prediction effectively, Jalali *et al.,* [11] advance structural design for online prediction in Web Usage Mining system and developed an innovative method based on LCS algorithm for classifying user navigation patterns for predicting users' future needs.

Web Usage Mining is one of the significant approaches for web recommendations, but the majority of its examinations are restricted in using web server log, and its applications are limited in serving a specific web site. In this approach, Yu Zhang *et al.,* [12] recommended a novel WWW-oriented web recommendation system based on mining the enterprise proxy log. The author initially evaluates the difference among the web server log and the enterprise proxy log, and then an incremental data cleaning approach is developed according to these differences. In data mining phase, this technique presented a clustering algorithm with hierarchical URL similarity. Experimental observation reveals that this system can implement the technology of Web Usage Mining effectively in this new field.

Data mining concentrates on the techniques of non-trivial extraction of inherent, previously unidentified, and potentially helpful information from extremely huge amount of data. Web mining is merely an application of data mining techniques to Web data. Web Usage Mining (WUM) is a significant class in Web mining. Web usage mining is an essential and rapid developing field of Web mining where numerous researches have been done previously. Jianxi Zhang *et al.,* [13] enhanced the fuzzy clustering approach to discover groups which share common interests and behaviors by examining the data collected in Web servers.

Web usage mining is one of the major applications of data mining techniques to logs of large Web data repositories with the aim of generating results used in some aspects, such as Web site design, user's classification, designing adaptive Web sites and Web site personalization. Data preprocessing is a vital phase in Web usage mining. The outcome of data preprocessing are significant to the next phases, like transaction identification, path examination, association rules mining, sequential patterns mining, etc. Zhang Huiying *et al.,* [14] used "USIA" algorithm was developed and its merits and demerits were examined, USIA is experimentally proved that not only its effectiveness is better and moreover it can recognize user and session accurately.

Web personalization systems are distinctive applications of Web usage mining. The Web personalization method is structured based on an online element and an off-line element. The off-line element is focused at constructing the knowledge base by examining past user profiles that is then utilized in the online element. Common Web personalization systems generally use offline data preprocessing and the mining procedure is not time-limited. On the other hand, this method is not a right choice in real-time dynamic environments. Consequently, there is a requirement for high-performance online Web usage mining approaches to offer solutions to these troubles. Chao *et al.,* [15] developed a comprehensive online data preprocessing process with the use of STPN. This approach developed the structural design for online Web usage mining in the data stream atmosphere and also developed an online Web usage mining system with the use of STPN that offers Web personalized online services.

3. Problems and Directions

Web usage mining helps in the prediction of interesting web pages in the website. Design assistance can be gathered from these data so as to increase its users. At the same time, the gathered data need to be consistent enough to predict the accurate data.

Several researchers proposed their ideas to enhance the web usage mining. The exiting works can be extended in order to satisfy the requirements in the following ways:

Initially, preprocessing can be improving by considering the addition information to remove the irrelevant web log records. This can be carried out by means of using the information such as browsing time, number of visits, etc.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Next, the focus is on grouping the browsing patterns. This will assists in better prediction. Therefore, the clustering algorithm used should be appropriate so as to perform better prediction. Also, in determining the user behaviors, the repeated sessions can be eliminated so as to avoid redundancy.

4. CONCLUSION

Web mining is the gathering of remarkable and helpful information and implicit data from the behavior of uses based on WWW, Web servers record and gathered data about user interactions every time demands for web pages are received. Examination of those Web access logs can assist in recognizing the user behavior and the web structure. When viewing from business and applications viewpoint, information gathered from the Web usage patterns can be directly utilized for efficiently manage activities corresponding to e-business, e-services, e-education, on-line communities, etc. Accurate Web usage data could assist to draw the attention of new customers, maintain present customers, enhances cross marketing/sales, effectiveness of promotional campaigns, track leaving customers and identifies the efficient logical structure for their Web space. User profiles could be constructed by merging users' navigation paths with other data characteristics like page viewing time, hyperlink structure, and page content. Conversely, as the size and complexity of the data escalated, the statistics suggested by conventional Web log examination techniques may prove insufficient and highly intelligent mining methods will be required. This paper discusses some of the existing web usage mining techniques and assist the researchers to develop a better strategy for web usage mining.

REFERENCES

[1] Heydari, M., Helal, R.A. and Ghauth, K.I., "A graph-based web usage mining method considering client side data", International Conference on Electrical Engineering and Informatics, Pp. 147-153, 2009.

[2] Qingtian Han, Xiaoyan Gao and Wenguo Wu, "Study on Web Mining Algorithm based on Usage Mining", 9th International Conference on Computer-Aided Industrial Design and Conceptual Design, Pp. 1121 – 1124, 2008.

[3] Inbarani, H.H., Thangavel, K and Pethalakshmi, A., "Rough Set Based Feature Selection for Web Usage Mining", International Conference on Computational Intelligence and Multimedia Applications, Pp. 33-38, 2007.

[4] Salin, S. and Senkul, P., "Using semantic information for web usage mining based recommendation", 24th International Symposium on Computer and Information Sciences, Pp. 236 – 241, 2009.

[5] Hussain, T., Asghar, S. and Fong, S., "A hierarchical cluster based preprocessing methodology for Web Usage Mining", 6th International Conference on Advanced Information Management and Service (IMS), Pp. 472-477, 2010.

[6] Gang Fang, Jia-Le Wang, Hong Ying and Jiang Xiong; "A Double Algorithm of Web Usage Mining Based on Sequence Number", International Conference on Information Engineering and Computer Science, 2009.

[7] Etminani, K., Delui, A.R., Yanehsari, N.R. and Rouhani, M., "Web usage mining: Discovery of the users' navigational patterns using SOM", First International Conference on Networked Digital Technologies, Pp. 224 – 249, 2009.

[8] Alphy Anna and Prabakaran, S., "Cluster optimization for improved web usage mining using ant nestmate approach", International Conference on Recent Trends in Information Technology (ICRTIT), Pp. 1271-1276, 2011.

[9] Yang Bin, Dong Xiangjun and Shi Fufu, "Research of WEB Usage Mining Based on Negative Association Rules", International Forum on Computer Science-Technology and Applications, Pp. 196-199, 2009.

[10] Aye, T.T., "Web log cleaning for mining of web usage patterns", 3rd International Conference on Computer Research and Development (ICCRD), Pp. 490 – 494, 2011.

[11] Jalali, M.; Mustapha, N.; Sulaiman, N.B.; Mamat, A., "A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems", 12th International Conference Information Visualisation, Pp. 302 – 307, 2008.

[12] Yu Zhang; Li Dai; Zhi-Jie Zhou, "A New Perspective of Web Usage Mining: Using Enterprise Proxy Log", International Conference on Web Information Systems and Mining (WISM), Pp. 38 – 42, 2010.

[13] Jianxi Zhang; Peiying Zhao; Lin Shang; Lunsheng Wang, "Web usage mining based on fuzzy clustering in identifying target group", International Colloquium on Computing, Communication, Control, and Management, Pp. 209 – 212, 2009.

[14] Zhang Huiying; Liang Wei, "An intelligent algorithm of data pre-processing in Web usage mining", Intelligent Control and Automation, Pp. 3119 – 3123, 2004.

[15] Chao, Ching-Ming; Yang, Shih-Yang; Chen, Po-Zung; Sun, Chu-Hao, "An Online Web Usage Mining System Using Stochastic Timed Petri Nets", 4th International Conference on Ubi-Media Computing (U-Media), Pp. 241 – 246, 2011.

[16] Hogo, M., Snorek, M. and Lingras, P., "Temporal Web usage mining", International Conference on Web Intelligence, Pp. 450-453, 2003.

[17] DeMin Dong, "Exploration on Web Usage Mining and its Application", International Workshop on Intelligent Systems and Applications, Pp. 1-4, 2009.

[18] Chih-Hung Wu, Yen-Liang Wu, Yuan-Ming Chang and Ming-Hung Hung, "Web Usage Mining on the Sequences of Clicking Patterns in a Grid Computing Environment", International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 6, Pp. 2909-2914, 2010.

[19] Tzekou, P., Stamou, S., Kozanidis, L. and Zotos, N., "Effective Site Customization Based on Web Semantics and Usage Mining", Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, Pp.51-59, 2007.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[20] Wu, K.L., Yu, P. S. and Ballman, A., "SpeedTracer: A Web usage mining and analysis tool", IBM Systems Journal, Vol. 37, No. 1, Pp. 89-105, 1998.

AUTHOR'S PROFILE

1. Ms. C. Thangamani

   Research Scholar

   Mother Terasa Women's University

   Kodaikanal.

2. Dr. P. Thangavel, Prof. & Head

   Department of Computer Science & Engineering

   Bannari Amman Institute of Technology

   Sathy.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A Comprehensive Comparison of the Performance of Fractional Coefficients of Image Transforms for Palm Print Recognition

Dr. H. B. Kekre
Sr. Professor,
MPSTME, SVKM's
NMIMS (Deemed-to-be
University, Vileparle(W),
Mumbai-56, India.

Dr. Tanuja K. Sarode
Asst. Professor
Thadomal Shahani Engg.
College,
Bandra (W), Mumbai-50,
India.

Aditya A. Tirodkar
B.E. (Comps) Student
Thadomal Shahani Engg.
College,
Bandra (W), Mumbai-50,
India.

Abstract

Image Transforms have the ability to compress images into forms that are much more conducive for the purpose of image recognition. Palm Print Recognition is an area where the usage of such techniques would be extremely conducive due to the prominence of important recognition characteristics such as ridges and lines. Our paper applies the Discrete Cosine Transform, the Eigen Vector Transform, the Haar Transform, the Slant Transform, the Hartley Transform, the Kekre Transform and the Walsh Transform on a two sets of 4000 Palm Print images and checks the accuracy of obtaining the correct match between both the sets. On obtaining Fractional Coefficients, it was found that for the D.C.T., Haar, Walsh and Eigen Transform the accuracy was over 94%. The Slant, Hartley and Kekre transform required a different processing of fractional coefficients and resulted with maximum accuracies of 88%, 94% and 89% respectively.

Keywords: Palm Print, Walsh, Haar, DCT, Hartley, Slant, Kekre, Eigen Vector, Image Transform

I. INTRODUCTION

Palm Print Recognition is slowly increasing in use as one highly effective technique in the field of Biometrics. One can attribute this to the fact that most Palm Print Recognition techniques have been obtained from tried and tested Fingerprint analysis methods [2]. The techniques generally involve testing on certain intrinsic patterns that are seen on the surface of the palm.

The palm prints are obtained using special Palm Print Capture Devices. The friction ridge impressions [3] obtained from these palm prints are then subjected to a number of tests related to identifying principal line, ridge, minutiae point, singular point and texture analysis [2][4][5][6]. The image obtained from the Capture devices however, is one that contains the entire hand and thus, software cropping methods are implemented in order to extract only the region of the hand that contains the palm print. This region, located on the hand's inner surface is called the Region of Interest (R.O.I.) [10][11][12][13]. Figure 1 shows us just how a Region of Interest is obtained from a friction ridge impression.

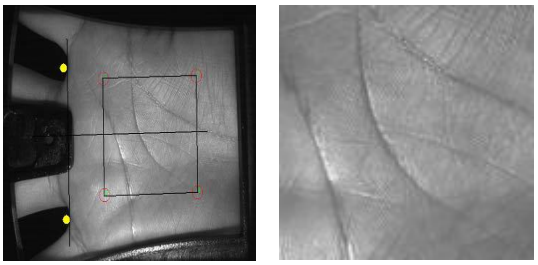


Fig.1 A on the left is a 2D-PalmPrint image from the Capture Device. B is the ROI image extricated from A and used for processing [3].

II. LITERATURE REVIEW

Palm Print Recognition like most Biometrics techniques constitutes the application of high performance algorithms over large databases of pre-existing images. Thus, it involves ensuring high accuracy over extremely large databanks and ensuring no dips in accuracy at the same time. Often, images with bad quality seem to ruin the accuracy of tests. Recognition techniques should also be robust enough to withstand such aberrations. As of now, literature based techniques involves the usage of obtaining the raw palm print data and subjecting it to transformations in order to transform it into a form that can be more easily used for recognition. This means that the data is to be arranged into feature vectors and then comparing called coding based techniques which are similar to those implemented in this paper. Other techniques include using line features in the palm print and appearance based techniques such as Linear Discriminant Analysis (L.D.A.) which are quicker but much less accurate techniques.

Transforms are coding models which are used on a wide scale in video/image processing. They are the discrete counterparts of continuous Fourier-related transforms. Every pixel in an image has a high amount of correlation that it shares with its neighbouring pixels. Thus, one can find out a great deal about a pixel's value if one checks this inherent correlation between a pixel and its surrounding pixels. By doing so, we can even correctly obtain the value of a pixel [1]. A transform is a paradigm that on application to such an image de-correlates the data. It does so by obtaining the correlation seen between a pixel and its neighbours and then concentrating the entropy of those pixels into one densely packed block of data. In most transformation techniques, we see that the data is found to be compressed into one or more particular corners. These areas that have a greater concentration of entropy can then be cropped out. Such cropped out portions are termed as fractional coefficients. It is seen that performing pattern

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

recognition on these cropped out images provides us with a much greater accuracy than with the entire image. Fractional Coefficients are generally obtained as given in Figure 2.

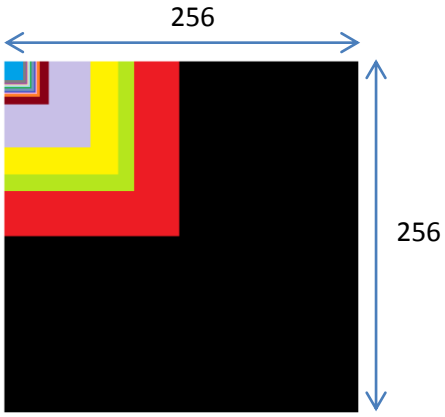


Figure 2. The coloured regions correspond to the fractional coefficients cropped from the original image, seen in black.

There are a number of such transforms that have been researched that provide us with these results. Some of them can be applied to Palm Print Recognition. In our paper, we apply a few of these transforms and check their accuracy for palm print recognition. The transforms we are using include the Discrete Cosine Transform, the P.C.A. Eigen Vector Transform, the Haar Transform, the Slant Transform, the Hartley Transform, the Kekre Transform and the Walsh Transform.

III. Implementation

Before we get to the actual implementation of the algorithm, let us see some pre-processing activities. Firstly, the database used consists of 8000 greyscale images of 128x128 resolution which contain the ROI of the palmprints of the right hand of 400 people. It was obtained from the Hong Kong Polytechnic University 2D_3D Database [7]. Here, each subject had ten palm prints taken initially. After an average time of one month, the same subject had to come and provide the palm prints again. Our testing set involved the first set of 4000 images from which query images were extracted and the second involved the next 4000. All these processing mechanisms were carried out in MATLAB R2010a. The total size of data structures and variables used totalled more than 1.07 GB.

One key technique that helped a great deal was the application of histogram equalization on the images in order to make the ridges and lines seem more prominent as seen in Figure 3. These characteristics are highly important as they form the backbone of most Palm Print Recognition technique parameters. In our findings, we have implicitly applied histogram equalization on all images. Without it, accuracy was found to be as low as 74% at average with most transforms. On the application of histogram equalization, it was found to increase to 94% in certain cases.

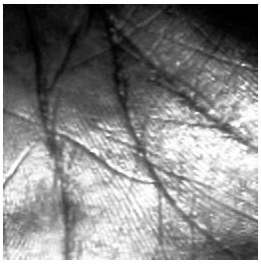


Figure 3. Histogram Equalized Image

IV. Algorithm

For our analysis, we carried out a set of operations on the databank mentioned above. The exact nature of these operations has been stated below in the form of an algorithm:

Step 1: Obtain the Query Image and perform Histogram Equalization on it.

Step 2: Apply the required Transformation on it.

Now, this image is to be compared against a training set of 4000 images. These images constitute the images in the database that were taken a month later.

Step 1: Obtain the Image Matrix for all images in the training set and perform Histogram Equalization on it.

Step 2: Apply the required Transform on each Image.

Step 3: Calculate the mean square error between each Image in the Training set and the query image. If partial energy coefficients are used, calculate the error between only that part of the images which falls inside the fractional coefficient. The image with the minimum mean square error is the closest match.

V. Transforms

Before providing the results of our study, first let us obtain a brief understanding of the plethora of transforms that are going to be applied in our study.

A. Discrete Cosine Transform

A discrete cosine Transform (DCT) is an extension of the fast Fourier Transform that works only in the real domain. It represents a sequence of finitely arranged data points in terms of cosine functions oscillating at different frequencies. It is of great use in compression and is often used to provide boundary functions for differential equations and are hence, used greatly in science and engineering. The DCT is found to be symmetric, orthogonal and separable [1].

B. Haar Transform

The Haar transform is the oldest and possibly the simplest wavelet basis. [9] [8]. Like the Fourier Analysis basis, it consists of square shaped functions which represents functions in the orthonormal function basis. A Haar Wavelet used both high-pass filtering and low-pass filtering and works by incorporating image decomposition on first he image rows and then the image columns. In essence, the Haar transform is one which when applied to

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

an image provides us with a representation of the frequency as well as the location of an image's pixels. It can thus be considered integral to the creation of the Discrete Wavelet Transforms.

C. Eigen Transform

The Eigen transform is a newer transform that is usually used as an integral component of Principal Component Analysis (P.C.A.). The Eigen Transform is unique as in it provides essentially a measure of roughness calculated from a pixels surrounding a particular pixel. The magnitude specified which each such measure provides us with details related to the frequency of the information [18][14]. All this helps us to obtain a clearer picture of the texture contained in an image. The Eigen transform is generally given by Equation 1:

$$Q(i,j) = \sqrt{\frac{2}{n+1}} \times \sin\frac{ij\pi}{n+1} \quad (1)$$

D. Walsh Transform

The Walsh Transform is a square matrix with dimensions in the power of 2. The entries of the matrix are either +1 or -1. The Walsh matrix has the property that the dot product of and two distinct rows or columns is zero. A Walsh Transform is derived from a Hadamard matrix of a corresponding order by first applying reversal permutation and then Gray Code permutation. The Walsh matrix is thus a version of the Hadamard transform that can be used much more efficiently in signal processing operations [19].

E. Hartley Transform

The Discrete Hartley Transform was first proposed by Robert Bracewell in 1983. It is an alternative to the Fourier Transform that is faster and has the ability to transform an image in the real domain into a transformed image that too stays in the real domain. Thus, it remedies the Fourier Transforms problem of converting real data into real and complex variants of it. A Hartley matrix is also its own inverse. For the Hartley Matrix we had to use a different method to calculate the fractional coefficients. This is because it polarizes the entropy of the image in all four corners instead of the one corner as seen with most transforms [15][16][17].

F. Kekre Transform

The Kekre Transform is the generic version of Kekre's LUV color space matrix. Unlike other matrix transforms, the Kekre transform does not require the matrix's order to be a power of 2. In the Kekre matrix, it is seen that all upper diagonal and diagonal elements are one while the lower diagonal elements below the sub diagonal are all zero. The diagonal elements are of the form –N+ (x-1) where N is the order of the matrix and x is the row coordinate [19]. The Kekre Transform essentially works as a high contrast matrix. Thus, results with the Kekre Transform are generally not as high as others. It too serves merely for experimental purposes.

G. Slant Transform

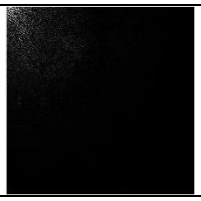
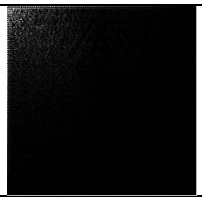
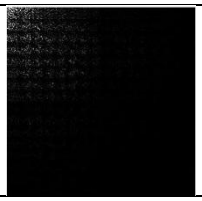
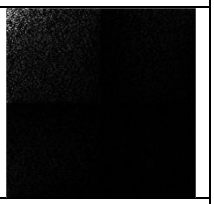
The Slant Transform is an orthonormal basis set of basis vectors specially designed for an efficient representation of those images that have uniform or approximately constant changing gray level coherence over a considerable distance of area. The Slant Transform basis can be considered to be a sawtooth waveform that changes uniformly with distance and represents a gradual increase of brightness. It satisfies the main aim of a transform to compact the image energy into as few of the transform components as possible. We have applied the Fast Slant Transform Algorithm to obtain it [20]. Like the Kekre, Hartley and Hadamard transforms, it too does not provide a good accuracy with the use of conventional fractional coefficient techniques [2]. For it, we have removed the fractional coefficient from the centre.

VI. Results

The results obtained for each transform with respect to their fractional coefficients are given in Table 1. Certain Transforms required a different calculation of fractional coefficients in order to optimize their accuracy. These transforms are given in Table 2 with their corresponding fractional coefficients.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE 1: COMPARISON TABLE OF ACCURACIES OBTAINED WITH DIFFERENT TRANSFORMS AT DIFFERENT FRACTIONAL COEFFICIENT RESOLUTIONS

	Accuracy			
Resolution	**D.C.T.**	**Eigen**	**Haar**	**Walsh**
Transformed Image				
256x256	92	92	92	92
128x128	91.675	91.8	91.7	92
64x64	93.3	93	93.425	93.525
40x40	94.05	93.65	93.675	94
32x32	94.3	94.075	93.925	94.175
28x28	94.225	94.2	94.05	94.3
26x26	94.275	94.35	**94.1**	**94.35**
25x25	94.375	94.4	94.025	94.25
22x22	94.4	94.325	93.95	94.025
20x20	**94.45**	94.425	94.025	93.95
19x19	94.4	**94.575**	93.7	93.85
18x18	94.425	94.5	93.6	93.8
16x16	94.25	94.375	93.375	93.675

From the above values, it is seen that for the purpose of Palm Print Recognition, all the above transforms viz. the Discrete Cosine Transform, the Eigen Vector Transform, the Haar Transform and the Walsh Transform are highly conducive and provide us with accuracy close to 94%. The highest accuracy is found in the case of the Eigen Vector transform with 94.575%. One factor of note is that all these maximum accuracies are obtained in a resolution range of 19x19 to 26x26 corresponding to fractional coefficients of 0.55% to 1.03%. Thus, in these cases, the processing required for operation is greatly decreased to a fraction of the original whilst providing an increase in accuracy. Let us see a comparison of the values in Table 1 with the help of the graph in Figure 4.

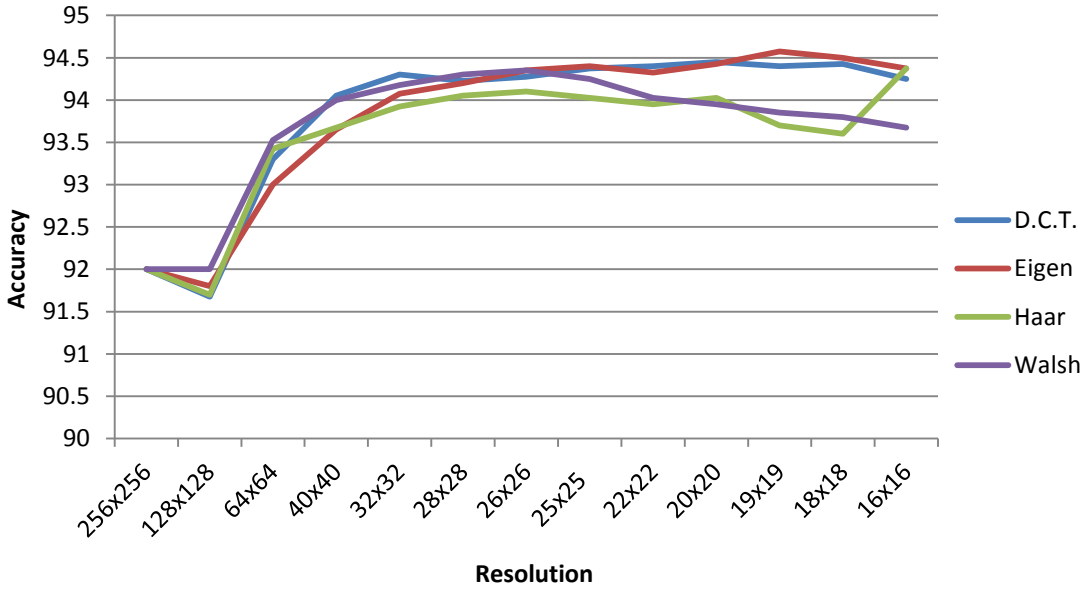
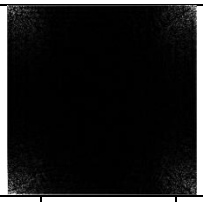
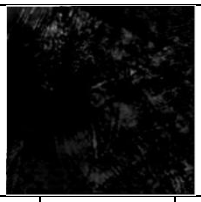
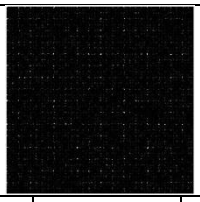


Figure 4: A Comparison Graph of Accuracy Values for the D.C.T., Eigen, Haar and Walsh Transforms.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE 2: ACCURACY COMPARISON FOR IMPROVISED FRACTIONAL COEFFICIENTS OF THE HARTLEY, KEKRE AND SLANT TRANSFORM

Hartley			**Kekre**			**Slant**		
								
Resolution	**Obtained From**	**Accuracy**	**Resolution**	**Obtained From**	**Accuracy**	**Resolution**	**Obtained From**	**Accuracy**
30x30	Matrices of order N/2 obtained from Each Corner	92.675	56x56	Selected From the Centre	72.25	128x128	Traditional	76.25
32x32		**94**	96x96		84.625	70x70	Selected From the Centre	83.075
62x62		93.025	127x127		88.975	80x80		81.575
128x128		92.5	128x128		**89.3**	128x128		**88.4**

Barring that of the Hartley matrix, in the above cases the accuracy of each transform is found to be much lower than that seen for the transforms tabulated in Table 1. This can be said because of the fact that these transforms do not polarize the energy values of the image pixels into any particular area of the image. The Hartley Transform requires all four corners to be considered, only then does it give us a good accuracy. The Kekre Transform as stated before works better as a high contrast matrix. When a Kekre contrasted matrix is subjected to a Discrete Cosine Transformation, it yields an accuracy of over 95%.

Thus, it can be termed as an intermediate transform, of more use in pre-processing than the actual recognition algorithm. The Slant Transform distributes the entropy across the entire image. This is highly cumbersome when it comes to calculating the mean square error. In all the above three algorithms, it is seen that obtaining the fractional coefficients requires some improvisation. With regular fractional coefficients, the above transforms yielded accuracies in the range of 70-75% with resolutions of 128x128.

VII. CONCLUSION

Thus, we can infer from our results that the D.C.T., Haar, Walsh and Eigen Vector Transforms yield credible accuracies of over 94% at fractional coefficients that lead to them providing a decrease in processing power roughly equal to 99% of that for the entire image. If the same method for obtaining fractional coefficients is used then for the Hartley, Kekre and Slant Transforms, we see a sharp decrease in accuracy. To amend this, improvisation is required as to obtaining the partial energy matrices. On doing so, we find the accuracy of the Hartley Matrix to increase to 94% that stands in league with the former four transforms. However, the accuracy in the case of the Slant and Kekre Transforms are still found to be less, providing maximum accuracy near 89%.

REFERENCES

[1] Syed Ali Khayam., "The Discrete Cosine Transform (DCT): Theory and Application." ECE 802-602: Information Theory and Coding. Seminar 1.

[2] Dr. H.B. Kekre, Sudeep D. Tepade, Ashish Varun, Nikhil Kamat, Arvind Viswanathan, Pratic Dhwoj. "Performance Comparison of Image Transforms for Palm Print Recognition with Fractional Coefficients of Transformed Palm Print Images." I.J.E.S.T. Vol.2(12), 2010, 7372-7379.

[3] Wei Li, Li Zhang, Guangming Lu, Jingqi Yan. "Efficient Joint 2D and 3D Palmprint Matching with Alignment Refinement." 23rd IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA. June 13-18, 2010.

[4] Parashar.S;Vardhan.A;Patvardhan.C;Kalra.P "Design and Implementation of a Robust Palm Biometrics Recognition and Verification System", Sixth Indian Conference on Computer Vision, Graphics & Image Processing.

[5]http://www.ccert.edu.cn/education/cissp/hism/039041.html (last referred on 29 Nov 2010)

[6] Kumar.A; Wong.D; Shen.H; Jain.A(2003): "Personal Verification Using Palm print and Hand Geometry Biometric." Proc. of 4th International Conference on Audio-and Video- Based Biometric Person Authentication (AVBPA)", Guildford, UK.

[7] PolyU 3D Palmprint Database, http://www.comp.polyu.edu.hk/~biometrics/2D_3D_Palmprint.htm

[8] Chin-Chen Chang, Jun-Chou Chuang and Yih-Shin Hu, 2004. "Similar Image Retrieval Based On Wavelet Transformation", International Journal Of Wavelets, Multiresolution And Information Processing, Vol. 2, No. 2, 2004, pp.111–120.

[9] Mohammed Alwakeel, Zyad Shaahban, "Face Recognition Based on Haar Wavelet Transform and Principal Component Analysis via Levenberg-Marquardt Backpropagation Neural Network."

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

European Journal of Scientific Research. ISSN 1450-216X Vol.42 No.1 (2010), pp.25-31

[10] W. Li, D. Zhang, L. Zhang, G. Lu, and J. Yan, "Three Dimensional Palmprint Recognition with Joint Line and Orientation Features", IEEE Transactions on Systems, Man, and Cybernetics, Part C, In Press.

[11] W. Li, L. Zhang, D. Zhang, G. Lu, and J. Yan, "Efficient Joint 2D and 3D Palmprint Matching with Alignment Refinement", in: Proc. CVPR 2010.

[12] D. Zhang, G. Lu, W. Li, L. Zhang, and N. Luo, "Palmprint Recognition Using 3-D Information", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Volume 39, Issue 5, pp. 505 - 519, Sept. 2009.

[13] W. Li, D. Zhang, and L. Zhang, "Three Dimensional Palmprint Recognition", IEEE International Conference on Systems, Man, and Cybernetics, 2009

[14] Tavakoli Targhi A., Hayman, Eric, Eklundh, Jan-Olof, Shahshanani, Mehrdad, "Eigen-Transform Transform Applications" Lecture Notes in Computer Science, 2006, Volume 3851/2006, 70-79

[15] John D. Villasenor "Optical Hartley Transform" Proceedings of the IEEE Vol. 82 No. 3 March 1994

[16] Vijay Kumar Sharma, Richa Agrawal, U. C. Pati, K. K. Mahapatra "2-D Separable Discrete Hartley Transform Architecture for Efficient FPGA Resource" Int'l Conf. on Computer & Communication Technology [ICCCT' 10]

[17] R.P. Millane "Analytic Properties of the Hartley Transform and their Implications" Proceedings of the IEEE, Col. 82, No. 3 March 1994

[18] Abdu Rahiman, V. Gigi C.V."Face Hallucination using Eigen Transformation in Transform Domain" International Journal of Image Processing (IJIP) Volume(3), Issue(6)

[19] Dr. H.B. Kekre, Dr. Tanuja K. Sarode, Sudeep D. Thepade, Sonal Shroff. "Instigation of Orthogonal Wavelet Transforms using Walsh, Cosine, Hartley, Kekre Transforms and their use in Image Compression." IJCSIS. Vol. 9. No. 6, 2011.Pgs. 125-133

[20] Anguh, Maurice, Martin, Ralph "A Truncation Method for Computing Slant Transforms with Applications to Image Processing" IEEE Transactions of Communications, Vol. 43, No. 6, June 1995.

AUTHOR BIOGRAPHIES

Dr. H. B. Kekre has received B.E. (Hons.) in Telecomm. Engineering. from Jabalpur University in 1958, M.Tech (Industrial Electronics) from IIT Bombay in 1960, M.S.Engg. (Electrical Engg.) from University of Ottawa in 1965 and Ph.D. (System Identification) from IIT Bombay in 1970 He has worked as Faculty of Electrical Engg. and then HOD Computer Science and Engg. at IIT Bombay. For 13 years he was working as a professor and head in the Department of Computer Engg. at Thadomal Shahani Engineering. College, Mumbai. Now he is Senior Professor at MPSTME, SVKM's NMIMS. He has guided 17 Ph.Ds, more than 100 M.E./M.Tech and several B.E./ B.Tech projects. His areas of interest are Digital Signal processing, Image Processing and Computer Networking. He has more than 300 papers in National / International Conferences and Journals to his credit. He was Senior Member of IEEE. Presently He is Fellow of IETE and Life Member of ISTE Recently seven students working under his guidance have received best paper awards. Currently 10 research scholars are pursuing Ph.D. program under his guidance.

Email: hbkekre@yahoo.com

Dr. Tanuja K. Sarode has Received Bsc.(Mathematics) from Mumbai University in 1996, Bsc.Tech.(Computer Technology) from Mumbai University in 1999, M.E. (Computer Engineering) degree from Mumbai University in 2004, Ph.D. from Mukesh Patel School of Technology, Management and Engineering, SVKM's NMIMS University, Vile-Parle (W), Mumbai, INDIA. She has more than 11 years of experience in teaching. Currently working as Assistant Professor in Dept. of Computer Engineering at Thadomal Shahani Engineering College, Mumbai. She is life member of IETE, member of International Association of Engineers (IAENG) and International Association of Computer Science and Information Technology (IACSIT), Singapore. Her areas of interest are Image Processing, Signal Processing and Computer Graphics. She has more than 100 papers in National /International Conferences/journal to her credit.

Email: tanuja_0123@yahoo.com

Aditya A. Tirodkar is currently pursuing his B.E. in Computer Engineering from Thadomal Shahani Engineering College, Mumbai. Having passionately developed a propensity for computers at a young age, he has made forays into website development and is currently pursuing further studies in Computer Science, looking to continue research work in the field of Biometrics.

Email: aditya_tirodkar@hotmail.com

Secured Dynamic Source Routing (SDSR) Protocol for Mobile Ad-hoc Networks

Dr. S. Santhosh Baboo
Reader, PG & Research Dept. of Computer Applications,
D.G.Vaishnav College,
Chennai, India
santhos2001@sify.com

S. Ramesh
Research Scholar,
Dravidian University,
Kuppam, Andra Pradesh, India
srameshdu@gmail.com

Abstract— *A mobile ad hoc network (MANET) is a collection of wireless mobile nodes dynamically shaping a provisional network devoid of the use of any existing network infrastructure or centralized management. In MANETs, security is the major challenge due to the dynamic topology which is because of the mobility of the nodes. In this paper, we propose to design and develop a secure methodology incorporated with the routing mechanism without having any compromise on the performance metrics viz., throughput, and packet delivery fraction. Not only just improving the throughput and packet delivery fraction it will also reduce the end-to-end delay and MAC overhead along with reduced packet loss. We name it as Secured-Dynamic Source Routing (SDSR) protocol. It adopts several features of the already existing protocol named Dynamic Source Routing (DSR). The simulation results prove that our proposed protocol SDSR outperforms DSR in all performance aspects.*

I. INTRODUCTION

The alluring infrastructure-less phenomenon of mobile ad hoc networks (MANETs) has received more attention in the research society. With the success of solving the most fundamental but vital issues in all network layers, persons understand there is commercial value in MANETs. The most of the applications that draw attention for utilizing in current wired networks (e.g., video conferencing, on-line live movies, and instant messenger with camera enabled) would attract interest for MANETs. Though, MANETs present distinctive advanced challenges, including the design of protocols for mobility management, effective routing, data transportation, security, power managing, and quality-of-service (QoS). Once these issues are resolved, the use of MANETs will be attainable. Nowadays applications heavily demand the fulfilment of their Quality of Service (QoS) requirements, which in this distributed and particular environment can be difficult to solve. This scenario requires specific proposals adapted to the new problem statements [3, 5, 12]. Trying to solve all these problems and coming out with a single solution would be too complex. To offer bandwidth-guaranteed QoS, the available end-to-end bandwidth along a route from the source to the destination must be known. The end-to-end throughput is a concave parameter [15], which is determined by the bottleneck bandwidth of the intermediate hosts in the route. A survey of several routing protocols and their performance comparisons have been reported in [4]. Hence in this paper, we focus on providing security along with QoS in MANETs.

In order to design good protocols for MANETs, it is important to understand the fundamental properties of these networks.

Dynamicity: Every node in the mobile ad hoc network will change its position on its own. Hence prediction of the topology is difficult, and the network status is not clear and it is vague.

Noncentralization: There is no existence of centralized control in mobile ad hoc network and, hence assigning resources to MANET in advance is not possible.

Radio properties: The medium is wireless, hence results in fading, multipath effects, time variation, etc. With these complications, Hard QoS is not easy to achieve.

II. RELATED WORKS

First, In [9] Zhao et al have reviewed the existing approaches of available bandwidth estimation. They presented the efforts and challenges in estimation of bandwidth. Also, they proposed a model for finding available bandwidth with improved accuracy of sensing based bandwidth estimation as well as prediction of available bandwidth.

In [17] Gui et al have defined routing optimality with the usage of different metrics like path length, energy consumption and energy aware load balancing within the hosts. Along with they have proposed a methodology for self-healing and optimizing routing (SHORT) technique for MANET. SHORT increases performance with regard to bandwidth and latency. They classified SHORT into two categories such as Path-Aware SHORT and Energy-Aware SHORT.

The QAMNet [14] approach extends existing ODMRP routing by introducing traffic prioritization, distributed resource probing and admission control mechanisms to provide QoS multicasting. For available bandwidth estimation, it used

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the same method given in SWAN [7] where the threshold rate for real-time flows is computed and the available bandwidth estimated as the deference between the threshold rate of real-time traffic and the current rate of real-time traffic. It is very difficult to estimate the threshold rate accurately because the threshold rate may change dynamically depending on traffic pattern [7]. The value of threshold rate should be chosen in a sensible way: Choosing a value that is too high results in a poor performance of real-time flows, and choosing a value that is too low results in the denial of real-time flows for which the available resource would have sufficed.

The localization methods are also distinguished by their form of computation, "centralized" or "decentralized". For example, MDS-MAP [6] is a centralized localization that calculates the relative positions of all the nodes based on connectivity information by Multidimensional Scaling (MDS). Similarly, DWMDS (Dynamic Weighted MDS) [11] uses movement constraints in addition to the connectivity information, and estimates the trajectories of mobile nodes. TRACKIE [13] first estimates mobile nodes that were likely to move between landmarks straight. Based on their estimated trajectories, it estimates the trajectories of the other nodes. Since these centralized algorithms use all the information about connectivity between nodes and compute the trajectories off-line, the estimation accuracy is usually better than decentralized methods.

In decentralized methods, the position of each node is computed by the node itself or cooperation with the other nodes. For example, APIT [16] assumes a set of triangles formed by landmarks, checks whether a node is located inside or outside of each triangle, and estimates its location. Amorphous [8] and REP [2] assume that location information is sent through multi-hop relay from landmarks, and each node estimates its positions based on hop counts from landmarks. In particular, REP first detects holes in an isotropic sensor network, and then estimates the distance between nodes accurately considering the holes. In MCL [15], each mobile node manages its Area of Presence (AoP) and refines its AoP whenever it encounters a landmark. In UPL [1], each mobile node estimates its AoP accurately based on AoP received from its neighboring nodes and obstacle information.

III. PROPOSED WORK

In order to implement QoS, we propose to develop a protocol which guarantees QoS along with secure dynamic source routing. In all the available existing protocols with regard to security, QoS requirements were compromised. We aim to develop a security enriched protocol which does not compromise with QoS requirements. For achieving the above goal we design a framework which uses estimation of 'bandwidth', estimation of 'residual energy', 'threshold value'.

A. Bandwidth Estimation

The bandwidth can be estimated as follows

Packet Delivery Time $(\mathbf{Ø_d}) = \mathbf{Ø_r - Ø_s}$

Where $Ø_r$ is Packet Received Time,

$Ø_s$ is Packet Sent Time

Bandwidth$= D_S / Ø_d \rightarrow (1)$

1

Where D_S is Data Size.

Bandwidth is the ratio between Size of the Data and Actual time taken to deliver the packet.

In following two cases Bandwidth gets reduced.

- When there is more channel contention i.e., Channel sensing busy due to more Request To Send (RTS) / Clear To Send (CTS) , collisions and higher backoffs.
- When there are more channel errors i.e., error bits in RTS/DATA which causes RTS/DATA retransmission.

B. Residual Energy

The Residual Energy [10] is calculated as follows:

$$RE_{node} = IE_{node} - CE_{node} \rightarrow 2$$

Where IEnode is the Initial Energy of the node and CEnode is the Consumed Energy of the node. The residual energy of a node is the difference between initial energy and consumed energy.

C. SDSR Routing

'Secured Dynamic Source Routing' (SDSR) is a routing protocol for MANETs. Our protocol SDSR uses distinct routing methodology. In which all the routing information is retained (updated again and again) at nodes. SDSR has only two foremost phases. They are Route Discovery and Route Maintenance. To identify source routes need collecting the address of each node from the source node to destination node in the course of route discovery. When the route discovery process is initiated, the two state-of-the art estimations such as bandwidth and residual energy will be calculated using (1) and (2). For making the reliable path, we have fixed the optimum bandwidth value to be 0.5 mbps. This optimum value will be suitable for the higher end applications like video-conferencing. The collected path information is cached by nodes which processes the route discovery packets. The path will be identified if the bandwidth is greater than or equal to 0.5 mbps so as to have more reliable path which assures QoS. The identified paths are used to route the packets. To achieve secured source routing, the routed packets will have the address of each node the packet will pass through. This may cause high overhead for longer paths in large scale mobile ad hoc network. To eliminate source routing, our SDSR protocol creates a stream id option which allows packets to be delivered based on a hop-by-hop basis.

Route Reply would only be produced when the message has reached the projected destination node. To send back the Route Reply, the destination node should have a route to the source node. The route would be used when the route is in the Destination Node's route cache. Or else, the node will turn round the route based on the route record in the Route Reply message header.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The Route Maintenance Phase will be started when there is an occurrence of incurable communication or when an Intruder node was identified using IDM. During above situation the Route Error packets are started at a node. The mistaken hop will be deleted from the node's route cache; all routes having the hop are terminated at that point. Once more, the Route Discovery Phase is started to find the most viable route.

D. *Intruder Detection Methodology (IDM)*

After calculating the path in which packets are to be routed, the source node will forward certain number packets to the next hop (node). The number of packets thus sent to the first hop will be set as threshold value. Thus obtained threshold value will be verified at every node in the path before despatching the packets. And if any of the node in the path has got different value other than that of threshold value then they are treated as Intruder and the path is rediscovered with the new threshold value and discarding the intruder node. Once again the above process is repeated till such time it reaches the destination node.

When the non-availability of a route to the next node, the node instantly updates the succession count and broadcasts the knowledge to its neighbors. When a node gets routing knowledge then it verifies in its routing table. If it does not have such entry into the routing table then updates the routing table with routing information it has obtained. If the node finds that it has already had an entry into its routing table then it compares the succession count of the received information with the routing table entry and updates the information. If it has succession count that is less than that of the received one then it rejects the information with the least succession count. Suppose both the succession counts are one and the same then the node keeps the information that has the shortest route or the least number of hops to that destination.

IV. PERFORMANCE METRICS

Average end-to-end delay: The end-to-end-delay is averaged over all surviving data packets from the sources to the destinations.

Average Packet Delivery Ratio: It is the ratio of the number of packets received successfully and the total number of packets sent.

Throughput: It is the number of packets received successfully.

Drop: It is the number of packets dropped.

V. RESULTS AND DISCUSSIONS

Figure 1 gives the throughput of both the protocols when the pause time is increased. As we can see from the figure, the throughput is more in the case of SDSR than DSR. Figure 2 presents the packet delivery ratio of both the protocols. Since the packet drop is less and the throughput is more, SDSR achieves good delivery ratio, compared to DSR. From Figure 3, we can ensure that the packets dropped are less for SDSR when compared to DSR. From Figure 4, we can see that the average end-to-end delay of the proposed SDSR protocol is less when compared to the DSR protocol.

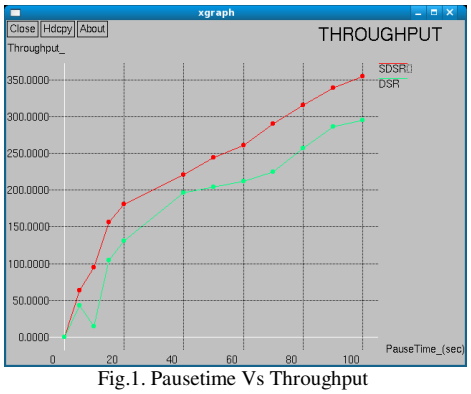


Fig.1. Pausetime Vs Throughput

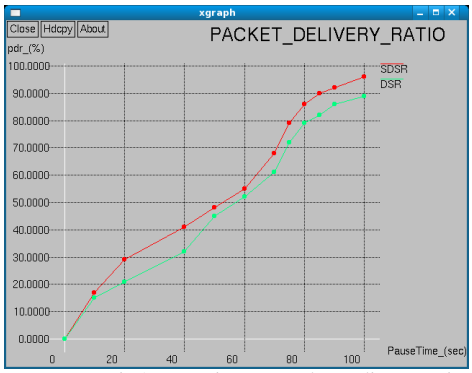


Fig.2. Pausetime Vs Packet Delivery Ratio

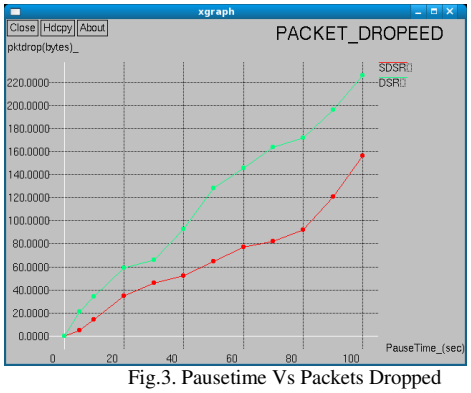


Fig.3. Pausetime Vs Packets Dropped

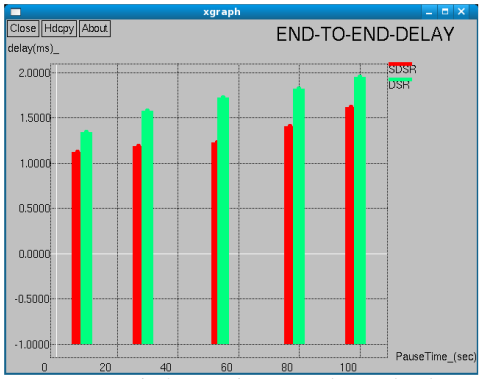


Fig.4. Pausetime Vs End-to-End Delay

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

VI. CONCLUSION AND FUTURE WORKS

In this paper we designed and developed a dynamic source routing named Secured Dynamic Source Routing (SDSR) protocol which meets the requirements of QoS such as improved throughput with better packet delivery ratio and reduced end-to-end delay and reduced no of drop in packets. Additionally, we provide a secure route maintenance mechanism by involving threshold in terms of packets. Further we provided security in terms of Advanced Encryption Standard (AES) algorithm using add-round key for data security while transmission of data. The results graph using the performance metrics outperformed when compared with Dynamic Source Routing (DSR) protocol. The framework used in this research would be further incorporated with other distance vector protocols.

REFERENCES

[1] A. Uchiyama, S. Fujii, K. Maeda, T. Umedu, H. Yamaguchi, and T. Higashino. Ad-hoc localization in urban district. In Proc. of INFOCOM 2007 Mini-Symposium, pages 2306–2310, 2007.

[2] M. Li and Y. Liu. Rendered path: range-free localization in anisotropic sensor networks with holes. In Proc. of MobiCom 2007, pages 51–62, 2007.

[3] Reddy T.B, Karthigeyan I, Manoj B. S, & Siva Ram Murthy C, "Quality of service provisioning in ad hoc wireless networks: A survey of issues and solutions", Ad hoc Networks, Vol. 4 No. 1, pp. 83–124, 2006.

[4] E.M. Royer and C.-K. Toh, "'A review of current routing protocols for ad hoc mobile wireless networks," in: IEEE Personal Communications,(April 1999).

[5] Chakrabarti S & Mishr A, "QoS issues in ad hoc wireless networks", IEEE Communications Magazine, Vol.39 No.2, pp.142–148, 2001.

[6] Y. Shang, W. Rml, Y. Zhang, and M. Fromherz. Localization from connectivity in sensor networks. IEEE Transaction on Parallel and Distributed Systems, 15(11):961–974, 2004.

[7] G. S. Ahn, A. T. Campbell, A. Veres and L.H. Sun, "SWAN: Service Differentiation in Stateless Wireless Ad hoc Networks", In Proc. IEEE INFOCOM, 2002.

[8] R. Nagpal, H. Shrobe, and J. Bachrach. Organizing a global coordinate system from local information on an ad hoc sensor network. In Proc. of IPSN 2003, pages 333–348, 2003.

[9] Haitao Zhao, Jibo Wei, Shan Wang and Yong Xi, "Available Bandwidth Estimation and Prediction in Ad hoc Networks", Wireless Networks, Vol.14, pp. 29–46, 2008.

[10] S. Santhosh Baboo, B. Narasimhan, "An Energy-Efficient Congestion-Aware Routing Protocol for Heterogeneous Mobile Ad Hoc Networks," act, pp.344-350, 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies, 2009.

[11] J. M. Cabero, F. D. la Torre, A. Sanchez, and I. Arizaga. Indoor people tracking based on dynamic weighted multidimensional scaling. In Proc. of MSWiM 2007, pages 328–335, 2007.

[12] Mohapatra & Gui C, "QoS in mobile ad hoc networks", IEEE Wireless Communications, Vol.10 No.3, pp. 44–52, 2003.

[13] S. Fujii, A. Uchiyama, T. Umedu, H. Yamaguchi, and T. Higashino. An off-line algorithm to estimate trajectories of mobile nodes using ad-hoc communication. In Proc. of PerCom 2008, pages 117–124, 2008.

[14] H. Tebbe, and A. Kassler, "QAMNet: Providing Quality of Service to Ad-hoc Multicast Enabled Networks", 1st International Symposium on Wireless Pervasive Computing (ISWPC), Thailand, 2006.

[15] P. Mohapatra, J. Li, and C. Gui, "QoS in mobile ad hoc networks," IEEE Wireless Commun. Mag. (Special Issue on QoS in Next-Generation Wireless Multimedia Communications Systems), pp. 44–52, 2003.

[16] T. He, C. Huang, B. M. Blum, J. A. Stankovic, and T. Abdelzaher. Range-free localization schemes for large scale sensor networks. In Proc. of MobiCom 2003, pages 81–95, 2003.

[17] Chao Gui & Mohapatra, "A Framework for Self-healing and Optimizing Routing Techniques for Mobile Ad hoc Networks", Wireless Networks, Vol.14 No.1, pp.29-46, 2008.

AUTHOR'S PROFILE

Lt.Dr.S.Santhosh Baboo, aged forty, has around Seventeen years of postgraduate teaching experience in Computer Science, which includes Six years of administrative experience. He is a member, board of studies, in several autonomous colleges, and designs the curriculum of undergraduate and postgraduate programmes. He is a consultant for starting new courses, setting up computer labs, and recruiting lecturers for many colleges. Equipped with a Masters degree in Computer Science and a Doctorate in Computer Science, he is a visiting faculty to IT companies. He has been keenly involved in organizing training programmes for students and faculty members. His good rapport with the IT companies has been instrumental in on/off campus interviews, and has helped the post graduate students to get real time projects. He has also guided many such live projects. Lt.Dr. Santhosh Baboo has authored a commendable number of research papers in international/ national Conference/ journals and also guides research scholars in Computer Science. Currently he is Reader in the Postgraduate and Research department of Computer Applications at Dwaraka Doss Goverdhan Doss Vaishnav College (accredited at 'A' grade by NAAC), one of the premier institutions in Chennai.

Ramesh Sadagoppan conceived his B.Sc.Chemistry and MCA degrees from University of Madras. He got his M.Phil Degree in Computer Science from Annamalai University. He is currently working as a Programmer in Centre for Railway Information Systems under Ministry of Railways in Chennai. He is currently pursuing his PhD Computer Science in Dravidian University under the research supervision of an eminent professor Lt.Dr.S.Santhosh Baboo.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Symbian 'vulnerability' and Mobile Threats

Wajeb Gharibi

Head of Computer Engineering &Networks Department, Computer Science & Information Systems College, Jazan University,
Jazan 82822-6694, Kingdom of Saudi Arabia
gharibi@jazanu.edu.sa

Abstract

Modern technologies are becoming ever more integrated with each other. Mobile phones are becoming increasing intelligent, and handsets are growing ever more like computers in functionality. We are entering a new era - the age of smart houses, global advanced networks which encompass a wide range of devices, all of them exchanging data with each other. Such trends clearly open new horizons to malicious users, and the potential threats are self evident.

In this paper, we study and discuss one of the most famous mobile operating systems 'Symbian'; its vulnerabilities and recommended protection technologies.

Keywords: *Information Security, Cyber Threats, Mobile Threats, Symbian Operating System.*

1. Introduction

Nowadays, there is a huge variety of cyber threats that can be quite dangerous not only for big companies but also for an ordinary user, who can be a potential victim for cybercriminals when using unsafe system for entering confidential data, such as login, password, credit card numbers, etc.

Modern technologies are becoming ever more integrated with each other. Mobile phones are becoming increasing intelligent, and handsets are growing ever more like computers in functionality. And smart devices, such as PDAs, on-board car computers, and new generation household appliances are now equipped with communications functions. We are entering a new era - the age of smart houses, global networks which encompass a wide range of devices, all of them exchanging data with each other via - as cyberpunk authors say - air saturated with bits and bytes. Such trends clearly open new horizons to malicious users, and the potential threats are self evident.

Our paper is organized as follows: Section 2 demonstrates the mobile operating system 'Symbian' vulnerabilities. Section3 proposes Symbians' Trojan Types. Section 4 recommends some possible protection techniques. Conclusions have been made in Section 5.

2. Symbian Vulnerabilities

The term 'vulnerability' is often mentioned in connection with computer security, in many different contexts. It is associated with some violation of a security policy. This may be due to weak security rules, or it may be that there is a problem within the software itself. In theory, all types of computer/mobile systems have vulnerabilities [1-5].

Symbian OS was originally developed by Symbian Ltd.[4]. It designed for smartphones and currently maintained by Nokia. The Symbian platform is the successor to Symbian OS and Nokia Series 60; unlike Symbian OS, which needed an additional user interface system, Symbian includes a user interface component based on S60 5th Edition. The latest version, Symbian^3, was officially released in Q4 2010, first used in the Nokia N8.

Devices based on Symbian accounted for 29.2% of worldwidesmartphone market share in 2011 Q1.[5] Some estimates indicate that the cumulative number of mobile devices shipped with the Symbian OS up to the end of Q2 2010 is 385 million [6].

On February 11, 2011, Nokia announced a partnership with Microsoft which would see it adoptWindows Phone 7 for smartphones, reducing the number of devices running Symbian over the coming two years.[12]

Symbian OS was subject to a variety of viruses, the best known of which is Cabir. Usually these send themselves from phone to phone by Bluetooth. So far, none have taken advantage of any flaws in Symbian OS – instead, they have all asked the user whether they would like to install the software, with somewhat prominent warnings that it can't be trusted.

This short history started in June 2004, when a group of professional virus writers known as 29A created the first virus for smartphones. The virus

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

called itself 'Caribe'. It was written for the Symbian operating system, and spread via Bluetooth. Kaspersky Lab classified the virus as Worm.SymbOS.Cabir.

Although a lot of media hype surrounded Worm.SymbOS.Cabir, it was actually a proof of concept virus, designed purely to demonstrate that malicious code could be created for Symbian. Authors of proof of concept code assert that they are motivated by curiosity and the desire to improve the security of whichever system their creation targets; they are therefore usually not interested either in spreading their code, or in using it maliciously. The first sample of Cabir was sent to antivirus companies at the request of its author. The source code of the worm was, however, published on the Internet, and this led to a large number of modifications being created. And because of this Cabir started too slowly but steadily infect telephones around the world.

A month after Cabir appeared, antivirus companies were startled by another technological innovation: Virus.WinCE.Duts. It occupies a double place of honour in virus collections - the first known virus for the Windows CE (Windows Mobile) platform, and also the first file infector for smartphones. Duts infects executable files in the device's root directory, but before doing this, requests permission from the user.

A month after Duts was born, Backdoor.WinCE.Brador made its appearance. As its name shows, this program was the first backdoor for mobile platforms. The malicious program opens a port on the victim device, opening the PDA or smartphone to access by a remote malicious user. Brador waits for the remote user to establish a connection with the compromised device.

With Brador, the activity of some of the most experienced in the field of mobile security - the authors of proof of concept viruses, who use radically new techniques in their viruses - comes almost to a standstill. Trojan.SymbOS.Mosquit, which appeared shortly after Brador, was presented as Mosquitos, a legitimate game for Symbian, but the code of the game had been altered. The modified version of the game sends SMS messages to telephone numbers coded into the body of the program. Consequently, it is classified as a Trojan as it sends messages without the knowledge or consent of the user - clear Trojan behaviour.

In November 2004, after a three month break, a new Symbian Trojan was placed on some internet forums dedicated to mobiles. Trojan.SymbOS.Skuller, which appeared to be a program offering new wallpaper and icons for Symbian was an SIS file - installer for Symbian platform. Launching and installing this program on the system led to the standard application icons (AIF files) being replaced by a single icon, a skull and crossbones. At the same time, the program would overwrite the original applications which would cease to function.

Trojan.SymbOS.Skuller demonstrated two unpleasant things about Symbian architecture to the world. Firstly, system applications can be overwritten. Secondly, Symbian lacks stability when presented with corrupted or non-standard system files - and there are no checks designed to compensate for this 'vulnerability'.

This 'vulnerability' was quickly exploited by those who write viruses to demonstrate their programming skills. Skuller was the first program in what is currently the biggest class of malicious programs for mobile phones. The program's functionality is extremely primitive, and created simply to exploit the peculiarity of Symbian mentioned above. If we compare this to PC viruses, in terms of damage caused and technical sophistication, viruses from this class are analogous to DOS file viruses which executed the command 'format c:\' .

The second Trojan of this class - Trojan.SymbOS.Locknut - appeared two months later. This program exploits the trust shown by the Symbian developers (the fact that Symbian does not check file integrity) in a more focused way. Once launched, the virus creates a folder called 'gavno' (an unfortunate name from a Russian speaker's point of view) in /system/apps. The folder contains files called 'gavno.app', 'gavno.rsc' and 'gavno_caption.rsc'. These files simply contain text, rather than the structure and code which would normally be found in these file formats. The .app extension makes the operating system believe that the file is executable. The system will freeze when trying to launch the application after reboot, making it impossible to turn on the smartphone.

3. Symbians' Trojan Types

Trojans exploiting the Symbian 'vulnerability' differ from each other only in the approach which is used to exploit the 'vulnerability'.

a) Trojan.SymbOS.Dampig overwrites system applications with corrupted ones

b) Trojan.SymbOS.Drever prevents some antivirus applications from starting automatically

c) Trojan.SymbOS.Fontal replaces system font files with others. Although the replacement files are valid, they do not correspond to the relevant language version of the font files of

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the operating system, and the result is that the telephone cannot be restarted

d) Trojan.SymbOS.Hoblle replaces the system application File Explorer with a damaged one

e) Trojan.SymbOS.Appdiasbaler and Trojan.SymbOS.Doombot are functionally identical to Trojan.SymbOS.Dampig (the second of these installs Worm.SymbOS.Comwar)

f) Trojan.SymbOS.Blankfont is practically identical to Trojan.SymbOS.Fontal

The stream of uniform Trojans was broken only by Worm.SymbOS.Lascon in January 2005. This worm is a distant relative of Worm.SymbOS.Cabir. It differs from its predecessor in that it can infect SIS files. And in March 2005 Worm.SymbOS.Comwar brought new functionality to the mobile malware arena - this was the first malicious program with the ability to propagate via MMS.

4. Possible Protection Techniques

Mobile has security vulnerabilities like computer and network. There is no particular locking system or guarding system that is able to ensure 100 percent security. Conversely, there are various types of security locks or guards that are suitable for different situations. We can make use of the combination of available and up to date technologies to fight the serious attacks. Yet there is no guaranty that this option will provide 100 percent security, nevertheless, this methodology certainly maximizes the mobile security and it is often possible to stop a threat. Few techniques are documented here which are also suggested by Wi-Fi Planet, 2007; TechRepublic, 2008; and TechGuru, 2010.

- Enable SIM, device and access lock from mobile settings. Enable the periodic lockdown feature. Enable the memory access code.
- Think deeply before accessing any internet site and installing any application.
- Spend little bit more time to check the application through Google or any search engine before downloading or installing unknown files.
- Disable WLAN and Bluetooth when you are out door and when you are not using it.
- Find a phone with the service option to remotely kill it when it is irretrievably lost.
- Never let others access your phone. Be careful while accepting calls or messages from unknown numbers.
- Enable WPA2 encryption for WLAN connection and pass code request feature for Bluetooth connection.
- If you noticed that your phone has connected to GPRS, UMTS, and HSDPA, disable those instantly.
- Keep regular backup.
- Install antivirus software.
- Do not simply save sensitive information on the phone unless absolutely essential.

5. Trends and forecasts

It is difficult to forecast the evolution of mobile viruses with any accuracy. This area is constantly in a state of instability. The number of factors which could potentially provoke serious information security threats is increasing more quickly than the environment - both technological and social - is adapting and evolving to meet these potential threats.

The following factors will lead to an increase in the number of malicious programs and to an increase in threats for smartphones overall:

- The percentage of smartphones in use is growing. The more popular the technology, the more profitable an attack will be.
- Given the above, the number of people who will have a vested interested in conducting an attack, and the ability to do so, will also increase.
- Smartphones are becoming more and more powerful and multifunctional, and beginning to squeeze PDAs out of the market. This will offer both viruses and virus writers more functionalities to exploit.
- An increase in device functionality naturally leads to an increase in the amount of information which is potentially interesting to a remote malicious user that isstored on the device. In contrast to standard mobile phones, which usually have little more than an address book stored on them, a smartphone memory can contain any files which would normally be stored on a computer hard disk. Programs which give access to password protected online services such as ICQ can also be used on smartphones, which places confidential data at risk.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

However, these negative factors are currently balanced out by factors which hinder the appearance of the threats mentioned above: the percentage of smartphones remains low, and no single operating system is currently showing dominance on the mobile device market. This currently acts as a brake on any potential global epidemic - in order to infect the majority of smartphones (and thus cause an epidemic) a virus would have to be multiplatform. Even then the majority of mobile network users would be secure as they would be using devices with standard (not smartphone) functionality.

Mobile devices will be under serious threat when the negative factors start to outweigh the positive. And this seems to be inevitable. According to data from the analytical group SmartMarketing, the market share of Symbian on the Russian PDA and smartphone market has been steadily increasing over the last 2 to 3 years. By the middle of 2005 it had a market share equal to that of Windows Mobile, giving rise to the possibility that the former may be squeezed out of the market.

Currently, there is no threat of a global epidemic caused by mobile malware. However, the threat may become real a couple of years down the line - this is approximately how long it will take for the number of smartphones, experienced virus writers and platform standardization to reach critical mass. Nevertheless, this does not reduce the potential threat - it's clear that the majority of virus writers are highly focussed on the mobile arena. This means that viruses for mobile devices will invariably continue to evolve, incorporating/ inventing new technologies and malicious payloads which will gradually become more and more widespread. The number of Trojans for Symbian which exploit the system's weak points will also continue to grow, although the majority of them are likely to be primitive (similar in functionality to Fontal and Appdisabler).

The overall movement of virus writers into the mobile arena is an equal stream of viruses analogous to those which are already known with the very rare inclusion of technological novelties and this trend seems likely to continue for the next 6 months at minimum. An additional stimulus for viruses writers will be the possibility of financial gain, and this will come when smartphones are widely used to conduct financial operations and for interaction with e-payment systems.

6. Conclusions

Smart mobile devices are still in their infancy, and consequently very vulnerable, both from a technical and a sociological point of view. On the one hand, their technical stability will improve only under arms race conditions, with a ceaseless stream of attacks and constant counter measures from the other side. This baptism of fire has only just begun for PDAs and smartphones, and consequently security for such devices is, as yet, almost totally undeveloped.

References

[1] Alexander Adamov, «Computer Threats: Methods of Detection and Analysis», Kaspersky Lab, Moscow 2009.

[2] www.securelist.com, «Examples and Descriptions of Various Common Vulnerabilities», Encyclopaedia.

[3] “Common Types of Mobile Malware” (2010) retrieved on 03rd April, 2010 from http://www.mobileantivirusstore.com/mobile-malware

[4] F-Secure “News From the Lab: Merogo SMS Worm” (2010) retrieved on 4th April, 2010 from http://www.fsecure.

[5] FortiGuard Center “Encyclopedia” (2010) retrieved on 10th April, 2010 from http://www.fortiguard.com/encyclopedia/virus/symbos_yxes.h!worm.html

[6] “Smartphones: Target for Hackers?” (2010) retrieved on 01st May, 2010 from http://pandalabs.pandasecurity.com/smartphones-target-for-hackers/

[7] Olzak, T. “Five Steps to Protect Mobile Devices Anywhere, Anytime” (2008) retrieved on 05th April, 2010 from http://blogs.techrepublic.com.com/security/?p=529

[8] Raywood, D. “Mobile Messaging Attacks to Rise in 2010” (2010) retrieved on 10th April, 2010 from http://www.securecomputing.net.au/News/165500,mobile-messaging-attacks-to-rise-in-2010.aspx

[9] “Nexus One” (2010) retrieved on 20th April, 2010 from http://www.google.com/phone/static/en_USnexusone_tech_specs.html

[10] “Mobile Threats” (2010) written by lecturer of Alluri Institute of Management Sciences, Warangal’ retrieved on 08Th May, 2010 from http://tricks9.info/2010/mobile-threats/

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Vertical Vs Horizontal Partition: In Depth

Tejaswini Apte
Sinhgad Institute of Business Administration and Research
Kondhwa(BK), Pune-411048
trapte@yahoo.com

Dr. Maya Ingle
Devi Ahilya VishwaVidyalay
Indore
maya_ingle@rediffmail.com

Dr. A.K.Goyal
Devi Ahilya VishwaVidyalay
Indore
goyalkcg@yahoo.com

***Abstract*-For the write-intensive operations and predictable behavior of queries, the traditional database system have optimize performance considerations. With the growing data in database and unpredictable nature of queries, write optimize system are proven to be poorly designed. Recently, the interest in architectures that optimize read performance by using Vertically Partitioned data representation has been renewed. In this paper, we identify the components affecting the performance of Horizontal and Vertical Partition, with the analysis. Our study focusing on tables with different data characteristics and complex queries. We show that carefully designed Vertical Partition may outperform carefully designed Horizontal Partition, sometimes by an order of magnitude.**

General Terms: Algorithms, Performance, Design

Keywords: Vertical Partition, Selectivity, Compression, Horizontal Partition

I. INTRODUCTION

Storing relational tables vertically on disk has been of keen interest as observed in data warehouse research community. The main reason lies in minimizing time required for disk reads for tremendously growing data warehouse. Vertical Partition (VP) possesses better cache management with less storage overhead. For queries retrieving more columns, VP demands stitching of the columns back together, offset the I/O benefits, potentially causing a longer response time than the same query on the Horizontal Partition (HP). HP stores tuples on physical blocks with slot array, specifies the offset of the tuple on the page [15]. HP approach is superior for queries, retrieve more columns and on transactional databases. For queries, retrieves less columns (DSS systems) HP approach may result in more I/O bandwidth, poor cache behavior and poor compressible ratio [6].

Current up-gradation of database technology has improved HP compression ratio by storing the tuples densely in the block, with poor updatable ratio and improved I/O bandwidth than VP. To achieve degree of HP compression close to entropy of table, skewed dataset and advance compression techniques opened the research path for response time of queries and HP performance for DSS systems [16].

Previous research shown results relevant to this paper are:

- HP is superior than VP, at less selectivity when query retrieves more columns with no chaining and the system is CPU constrained.
- Selectivity factor and number of retrieved columns is the measure of processing time of VP than HP.
- VP may be sensitive to the amount of processing needed to decompress a column.

Compression ratio may be improved for non-uniform distribution [13]. Research community mainly focused on single predicate with less selectivity, applied to the first column of the table, and the same is retrieved by the query [12]. We believe that the relative performance of VP and HP is affected by (a) Number of Predicates (b) Predicates application on columns and Selectivity (c) Resultant Columns. Our approach mainly focusing on factors, affecting response time of HP and VP i.e. (a) Additional Predicate (b) Data Distribution (c) Join Operation.

For various applications, it has been observed that VP has several advantages over HP. We discuss related, existing and recent compression techniques of HP and VP in Section 2. Many factors affects the performance of HP and VP. Section 3 provides the comparative study of performance measure with query characteristics. Our approach's implementation detail and analysis of the result is presented in Section 4. Finally, we conclude with a short discussion of our work in Section 5.

II. RELATED WORK

In this section, some existing compression techniques used in VP and HP have been discussed briefly along with the latest methodologies.

A. Vertical Storage

The VP and HP comparison is presented with C-Store and Star Schema Benchmark [12]. VP is implemented using commercial relational database systems by making each column its own table. The idea presented had to pay more performance penalty, since every column must have its own row-id. To prove the superiority of HP over VP, analysis has done by implementing HP in C-store (VP database). Compression, late materialization and block iteration were the base of measure for the performance of VP over HP.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

With the given workload, compression and late materialization improves performance by a factor of two and three respectively [12]. We believe these results are largely orthogonal to ours, since we heavily compress both the HP and VP and our workload does not lend itself to late materialization of tuples. "Comparison of Row Stores and Column Stores in a Common Framework" mainly focused on super-tuple and column abstraction. Slotted page format in HP results in less compression ratio than VP [10]. Super-tuples may improve the compression ratio by storing rows with one header with no slot-array. Column abstraction avoids storing repeated attributes multiple times by adding information to the header. Comparison is made over varying number of columns with uniformly distributed data for VP and HP, while retrieving all columns from table.

The VP concept has implemented in Decomposition storage model (DSM), with storage design of (tuple id, attribute values) for each column (MonetDB) [9]. C-Store data model contains overlapping projections of tables. L2 cache behaviour may improved by PAX architecture, focused on storing tuples column-wise on each slot [7], with penalty of I/O bandwidth. Data Morphing improves on PAX to give even better cache performance by dynamically adapting attribute groupings on the page [11].

B. Database Compression Techniques

Compression techniques in database is mostly based on slotted page HP. Compression ratio may be improved up to 8-12 by using processing intensive techniques [13]. VP compression ratio is examined by "Superscalar RAM-CPU Cache Compression" and "Integrating Compression and Execution in Column-Oriented Database Systems" [21, 3]. Zukowski presented an algorithm for compression optimization the usability of modern processor with less I/O bandwidth. Effect of run lengths on degree of compression and dictionary encoding proven to be best compression scheme for VP [3].

III. PERFORMANCE MEASURING FACTORS

Our contribution to existing approach is based on the major factors affecting the performance of HP and VP (a)Data Distribution (b)Cardinality (c)Number of columns (d)Compression Technique and (e) Query nature.

A. Data Characteristics

The search time, and performance of two relational tables varies with number of attributes, data type of each attribute along with the compression ratio, column cardinality and selectivity.

B. Compression Techniques

Dictionary based coding

The repeated occurrences are replaced by a codeword that points to the index of the dictionary that contains the pattern. Both code words and uncompressed instructions are part of compressed program. Performance penalty occurs for (a) Dictionary cache line is bigger than processors L1 data cache (b) Index size is larger than value and (c) Un-encoded column size is smaller than the size of the encoded column plus the size of the dictionary [3].

Delta coding

The data is stored, as the difference between successive samples (or characters). The first value in the delta encoded file is the same as the first value in the original data. All the following values in the encoded file are equal to the difference (delta) between the corresponding value in the input file, and the previous value in the input file. For uniform values in the database, delta encoding for data compression is beneficial. Delta coding may be performed on both column level and tuple level. For unsorted sequence and size-of(encoded) is larger than size-of(un-encoded), delta encoding is less beneficial [3].

Run Length Encoding (RLE)

The sequences of the same data values within a file is replaced by a count number and a single value. RLE compression works best for sorted sequence, long runs. RLE is more beneficial for VP [3].

C. Query Parameters and Table Generation

To study the effect of queries with table characteristics, queries were tested with varying number of predicates and selectivity factor. Factors affecting the execution plan and cost are (a)Schema definition (b) Selectivity factor (c) Number of columns referenced (d) Number of predicates. The execution time of a query change with column characteristics and I/O bandwidth. For each characteristic of column, the query generator randomly selects the columns used to produce a set of "equivalent" queries with the cost analysis [12]. Performance measure with compression is implemented by:

- Generation of uncompressed HP version of each table with primary key on left most column.
- Sorted on columns frequently used in query.
- Replica is generated on VP.

IV. IMPLEMENTATION DETAIL

To study the effect of VP and HP, the experiments are done against TPC-H standard Star-Schema on MonetDB.

We mainly concentrated on the fact table i.e. Sales, contains approximately 10L records. We focused on five columns for selectivity i.e. prod_id, cust_id, time_id, channel_id, promo_id with selectivity varying from 0.1 to 50%.

```
SELECT p.product_name,ch.channel_class,
c.cust_city, t.calendar_quarter_desc,
  SUM(s.amount_sold) sales_amount
FROM sales s, times t, customers c, channels ch,
products p, promotions pr
WHERE s.time_id = t.time_id
AND s.prod_id=p.prod_id
AND   s.cust_id = c.cust_id
AND   s.channel_id = ch.channel_id
```


http://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
AND s.promo_id=pr.promo_id
AND   c.cust_state_province = 'CA'
AND   ch.channel_desc in ('Internet','Catalog')
AND   t.calendar_quarter_desc IN ('1999-Q1','1999-
Q2')
GROUP BY ch.channel_class,p.product_name
c.cust_city, t.calendar_quarter_desc;
```

Table 1: Generalized Star-Schema Query

A. *Read-Optimized Blocks (Pages)*

The HP and VP, dense pack the table on the blocks to achieve less I/O bandwidth. With varying page size HP keeps tuples together, while the VP stores each column in a different file. The different entries on the page are not aligned to byte or word boundaries in order to achieve better compression. Each page begins with the page header, contains number of entries on the page, followed by data and compression dictionary. The size of the compression dictionary is stored at the very end of the page, with the dictionary growing backwards from the end of the page towards the front. For the HP, the dictionaries for the dictionary-compressed columns are stored sequentially at the end of the page.

B. *Query Engine, Scanners and I/O*

The query scanner scans the files differently for HP and VP. Materialization of results are done after reading the data and applying predicates to it, with minimum passes in HP than VP, which requires reading multiple files for each column referenced by query. Predicates are applied on a per-column basis, columns are processed by order of their selectivity, most selective (with the fewest qualifying tuples) to least selective (the most qualifying tuples). Placing the most selective predicate first allows the scanner to read more of the current file before having to switch to another file, since the output buffer fills up more slowly.

C. *Experimental Setup*

All results were run on a machine running RHEL 5 on a 2.4 GHz Intel processor and 1GB of RAM. HP and VP are affected by the amount of I/O and processing bandwidth available in the system; for each combination of output selectivity and number of columns accessed.

Effect of selectivity

Selecting fewer tuples with very selective filter and index has no effect on I/O performance, system time remains the same. The HP remains the same, since it has to examine each tuple in the relation to evaluate the predicate. For the VP evaluating the predicate requires more time. With decrease in selectivity VP and HP performance ratio is less. However as selectivity increases towards 100%, each column scan contribute in CPU cost. The VP is faster than HP when more columns are returned with the selectivity factor from 0.1% to 25%. Further with same configuration compressed HP will speed up by 4 in VP (Figure 1).

Predicate	Selectivity(%)	No Of Rows	HP(time in sec)	VP(time in sec)
Prod_id	Compressed(50)	1000000	3	14
Cust_id	25	1000000	45	10
Time_id	10	10,00000	40	20
Promo_id	1	1000000	35	20
Channel_id	0.1	1000000	30	30

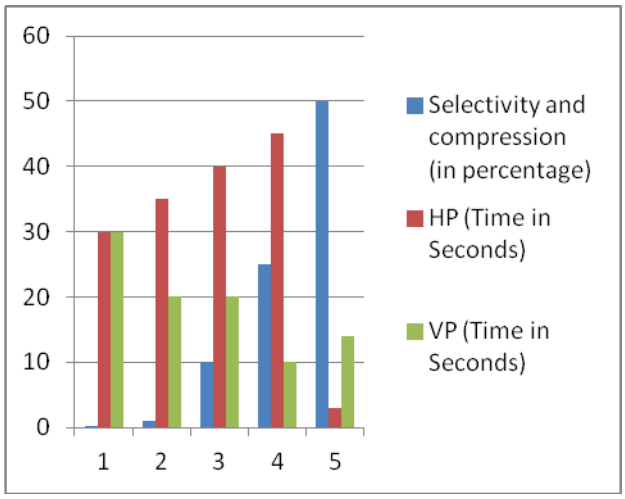


Figure 1: Time measurement for HP and VP with varying selectivity and Compression

Effect of compression

For skew data distribution and large cardinality in HP, run-length and dictionary compression techniques are more beneficial. The size of VP tuple is approximately same as size of HP tuple. HP compression is a critical component in determining its performance relative to that of the VP. Compression is more beneficial for columns having high cardinality. For compression, some VP proponents have argued that, since VP compress better than HP, storing the data with multiple projections and sort orders are feasible and can provide even better speedups [18].

Effect of Joins

We examined join operations for query presented in table 1, with varying predicates over HP and VP, to analyze the interaction of resultant tuple with join (e.g. more instruction cache misses due to switching between scanning and reconstructing tuples and performing the join).

Compression improves the performance by decreasing I/O bandwidth and increasing scan time, as the columns selection ratio grows. Unlike compression, cost of join operation has increased with increased list of selected columns. The HP outperforms the VP as number of accessed columns is more. The join component of the time is always roughly equivalent between the HP and VP (Figure 2). Thus, the paradigm with the smaller scan time will also have the smaller join time, and

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the join time is greatly affected by the number of joined tuples materialized, number of passes are required, the type of join operation.

Join Operations	VP (time in sec)	HP (time in sec)
Hash Join	27	30
Merge Join	38	35
Nested Loop Join	30	34

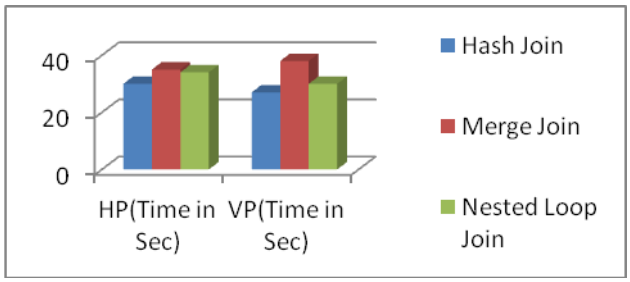


Figure 2: Performance of Join Operation in HP and VP

Analysis

Our analysis focuses tuple-at-a-time paradigm. The cost for each tuple evaluation is the minimum of CPU processing and Disk bandwidth. Performance of the database depends on size of input (SOI). For any query,

Total Disk Rate (TDR) = SOI1/TOS+ … +SOI n/TOS

For more columns, HP outperforms the VP. CPU cost measured by search and operations time on the query.
Thus it is,
*Cost (*CPU) = Cost (Operations)||Cost(Scan)
Rate of an operator
OP=time/no. of CPU instructions

V. CONCLUSION

We summaries the following points:

A. The selectivity of predicate can substantially change the relative performance of HP and VP.
B. HP performs better compared to VP, when most of the columns are required by the query.
C. Adding predicates increases VP run times.
D. Joins do not change the relative performance of HP and VP.

VP outperforms a HP when I/O is a dominating factor in query plan and for less columns selection. For HP with compression, I/O becomes less of a factor and CPU time is more of a factor in VP for queries with more predicates, lower selectivity and more columns referenced. HP on slotted pages will most likely never beat VP for read-optimized workloads.

REFERENCES

[1]https://www.cs.hut.fi/Opinnot/T106.290/K2005/Ohjeet/Zipf.html. Accessed November 8,2007.
[2] http://db.lcs.mit.edu/projects/cstore/. Accessed November 8, 2007.
[3] Abadi, D. J., Madden, S. R., Ferreira, M. C. "Integrating Compression and Execution in Column-Oriented Database Systems." In *SIGMOD*, 2006.
[4] Abadi, D.J., Madeen, S. R., Hachem, N. "Column-Stores vs. Row-Stores: How Different Are They Really?" In *SIGMOD*, 2008.
[5] Abadi, D. J., Myers, D.S., DeWitt, D.J., Madden, S.R. "Materialization Strategies in a Column-Oriented DBMS." In *ICDE*, 2007.
[6] Ailamaki, A. *Architecture-Conscious Database Systems*. Ph.D. Thesis, University of Wisconsin, Madison, WI, 2000.
[7] Ailamaki, A., DeWitt, D. J., Hill, M. D., and Skounakis, M. "Weaving Relations for Cache Performance." In *VLDB*, 2001.
[8] Boncz, P., Zukowski, M., and Nes, N. "MonetDB/X100: Hyper-Pipelining Query Execution." In *CIDR*, 2005.
[9] Copeland, A. and Khoshafian, S. "A Decomposition Storage Model." In *SIGMOD*, 1985.
[10] Halverson, A. J., Beckmann, J. L., Naughton, J. F., DeWitt, D. J. "A Comparison of C-Store and Row-Store in a Common Framework." *Technical Report, University of Wisconsin-Madison, Department of Computer Sciences*, T1666, 2006.
[11] Hankins, R. A., Patel, J. M. "Data Morphing: An Adaptive, Cache-Conscious Storage Technique." In *VLDB*, 2003.
[12] Harizopoulos, S., Liang, V., Abadi, D., and Madden, S. "Performance Tradeoffs in Read-Optimized Databases." In *VLDB*, 2006.
[13] Holloway, A. L., Raman, V., Swart, G. and DeWitt, D. J. "How to Barter Bits for Chronons: Compression and Bandwidth Trade Offs for Database Scans." In *SIGMOD*, 2007.
[14] Huffman, D. "A Method for the Construction of Minimum-Redundancy Codes." In *Proceedings of the I. R. E.*, pages 1098-1102, 1952.
[15] Ramakrishnan, R. and Gehrke, J. *Database Management Systems*. McGraw-Hill, 3rd edition, 2003.
[16] Raman, V., Swart, G. "Entropy Compression of Relations and Querying of Compressed Relations." In *VLDB*, 2006.
[17] Shapiro, L. D. "Join processing in database systems with large main memories." *ACM Trans. Database Syst*. 11(3): 239-264 (1986).
[18] Stonebraker, M., et al. "C-Store: A Column-Oriented DBMS." In *VLDB*, 2005.
[19] T. P. P. Council. "TPC Benchmark H (Decision Support)," http://www.tpc.org/tpch/default.asp, August 2003.
[20] "The Vertica Database Technical Overview White Paper." Vertica, 2007.
[21] Zukowski, M., Heman, S., Nes, N., and Boncz, P. "Super-Scalar RAM-CPU Cache Compression." In *ICDE*, 2006.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Framework for Query optimization

Pawan Meena
Department of
Computer Science and Engineering
Patel college of science & Technology
Bhopal,M.P,INDIA
pawanmeena75@yahoo.com

Arun Jhapate
Department of
Computer Science and Engineering
Patel college of science & Technology
Bhopal,M.P,INDIA
Arun_jhapate@yahoo.com

Parmalik kumar
Department of
Computer Science and Engineering
Patel college of science & Technology
Bhopal,M.P,INDIA
Parmalik83@gmail.com

ABSTRACT

Modern database systems use a query optimizer to identify the most efficient strategy, called "plan", to execute declarative SQL queries. Optimization is much more than transformations and query equivalence. The infrastructure for optimization is significant. Designing effective and correct SQL transformations is hard. Optimization is a mandatory exercise since the difference between the cost of the best plan and a random choice could be in orders of magnitude. The role of query optimizers is especially critical for the decision-support queries featured in data warehousing and data mining applications. This paper presented an abstraction of the architecture of a query optimizer and focused on the techniques currently used by most commercial systems for its various modules. In aaddition, provide technical constraint of advanced issues in query optimization.

Keywords

Query optimizer ,Operator tree, Query analyzer, Query optimization

1. Introduction

For significantly improve application development and user productivity, relational database technology growing success in the treatment of data is appropriate in part to the availability of non-procedural languages. By hiding the low-level details about the physical organization of the data, relational database languages allow the expression of complex queries in a concise and simple fashion. In particular, to build the answer to the query, the user does not exactly specify the procedure. This procedure is in fact designed by a DBMS module, known as query processor. This relieves the user to query optimization, a tedious task that is managed correctly by the query processor. Modern databases can provide tools for the effective treatment of large amounts of complex scientific data involving the application of specific analysis [1, 2]. Scientific analysis can be specified as high-level requests user-defined functions (UDFs) in an extensible DBMS. The query optimization provides scalability and high performance without the need for researchers to spend time on low-level programming. Moreover, as the queries are specified and easily changed, new theories, for example implemented as filters, can be tested quickly.

Queries about events are complex, because the cuts are complex with many predicates applied to the properties of each event. The conditions of the query involving selections, arithmetic operators, aggregates, UDF, and joins. The aggregates compute complex derived event properties. For example, a complex query is to look for event production Higgs bosons [1, 3] by applying scientific theories expressed cuts. These complex queries need to be optimized for the efficient and scalable. However, the optimization of complex queries is a challenge because:

• The queries contain many joins.

• The size of the queries makes optimization slow.

• The cut definitions contain many more or less complex aggregates.

• The filters defining the cuts use many numerical UDFs.

• There are dependencies between event properties that are difficult to find or model.

• The UDFs cause dependencies between query variables.

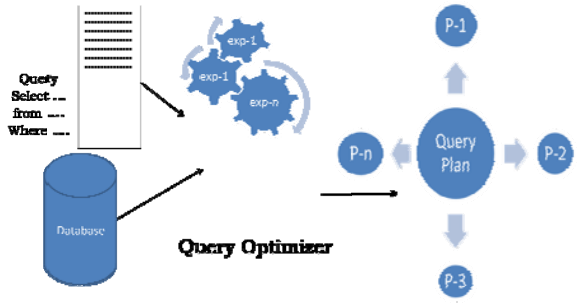


Figure 1: Query Optimizer

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Relational query languages provide a high level "declarative" interface to access data stored in relational databases. Over time, SQL [1,4] has emerged as the standard for relational query languages. Two key elements of the component of the evaluation of a system for querying SQL databases are the query optimizer and execution engine queries. The query execution engine implements a set of physical operators. An operator takes as input one or more data streams and produces an output data stream. Examples of operators are physical (external) sorting, sequential analysis, index analysis, nested loop join and sort-merge join. We refer to operators such as physical operators since they are not necessarily related one by one with the relational operators. The easiest way to think of physical operators is like pieces of code that are used as building blocks to enable the execution of SQL queries. An abstract representation of such a performance is a physical operator tree, as shown in Figure 2. The edges in an operator tree represent the flow of data between the physical operators.

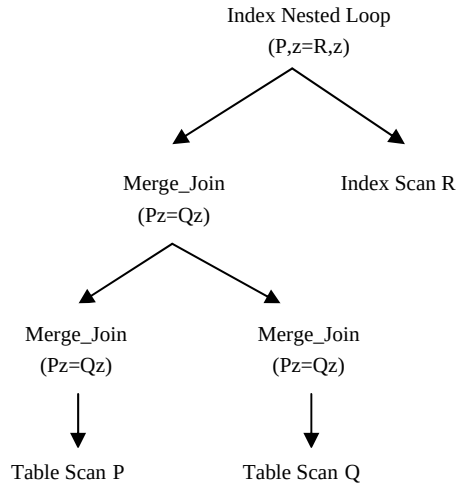


Figure 2: Physical Operator Tree

We use the terms physical operator tree and execution plan (or simply plan) interchangeably. The execution engine is responsible for implementing the plan resulting generate responses to the request. Therefore, the Capabilities of the query execution engine to determine the structure of the operator trees that are practicable. We refer the reader to [5] for an overview of the technical evaluation of the query. The query optimizer is responsible for producing the input for the execution engine. It takes a parsed representation of an SQL query as input and is responsible for producing an efficient execution plan for the given SQL query in the space of possible execution plans. The task of an optimizer is nontrivial since for a given SQL query, there may be many operator trees possible:

• The algebraic representation of the data query can be transformed into many other logically equivalent algebraic representations: for example,

Join (Join (P, Q), R) = Join (Join (Q, R), P)

• For a given algebra representation, there can be many operator trees that the operator algebraic expression to perform, for example, in general, there are several algorithms supported them in a system database. In addition, the current or the response time for the implementation of these plans is very different. Therefore, a choice of execution by the optimization program is crucial. For instance, query optimizations are regarded as difficult search. To solve this problem, we need:

• A space of plans (search space).

• A cost estimation technique so that a cost may be assigned to each plan in the search space. Intuitively, this is an estimation of the resources needed for the execution of the plan.

• An enumeration algorithm that can search through the execution space A desirable optimizer is one where the search space includes plans to lower costs, the costing technique is correct and the enumeration algorithm efficient. Each of these tasks is nontrivial and that is why building a good optimizer is a huge undertaking.

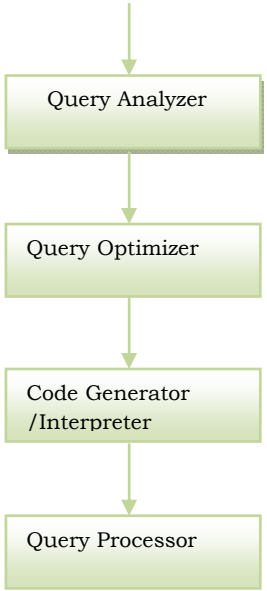


Figure 3: Query traverses through DBMS

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The path through a query to a DBMS is generated by its reaction is shown in Figure 3.The modules of the system, allowing it to move the following functions.

The Query Analyzer checks the validity of the query; it creates an internal form, usually an expression of the relational calculus or something similar. The query optimizer considers all algebraic expressions that are equivalent to the given query and choose one that is estimated to be less expensive. The code generator or interpreter changes the map generated by the optimizer calls the query processor.

2. Query Optimization Architecture

In this section, we provide an abstraction of the query optimization process in a DBMS. Given a database and a query on it, several execution plans exist that can be employed to answer the query. In principle, all the alternatives need to be considered so that the one with the best estimated performance is chosen. An abstraction of the process of generating and testing these alternatives is shown in Figure 4, which is essentially a modular architecture of a query optimizer. Although one could build an optimizer based on this architecture, in real systems, the modules shown do not always have so clear-cut boundaries as in Figure 4. Based on Figure 4, the entire query optimization process can be seen as having two stages: rewriting and planning [6]. There is only one module in the first stage, the Rewriter, whereas all other modules are in the second stage. The functionality of each of the modules in Figure 4 is analyzed below

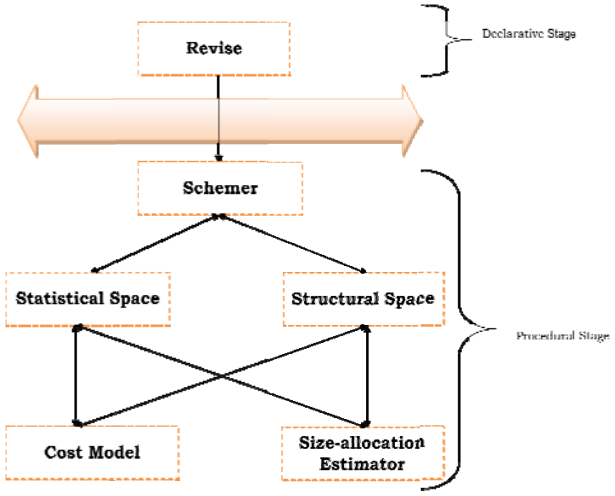


Figure 4: Query optimizer architecture

Revise: This module applies transformations to a given query and produces similar questions that are hopefully more effective, for example, replacement of thought with their definition, to attend nested queries, etc. The processing is done by the author only on the declarative, that is, static the characteristics of requests and do not take into account the actual cost for the specific question DBMS and the database in question. If rewriting is known or assumed always positive, the initial request is ignored, otherwise sent to the next as well. The nature of the transformations to rewrite this step occurs in declarative level [6].

Schemer: This is the main module of the ordering stage. Examine all possible execution plans for each query generated in the previous step and selects the best global market to be used for the reaction to generate the original query. It employs a research strategy that examines the space of execution plans in a particular fashion. This is determined by two other modules of the optimizer, space and space-mode algebraic structure. Most of these modules and the search strategy to the cost, i.e., work time, the optimizer itself, which should be as low as possible to determine. The implementations of the plans reviewed by the planner are compared in terms of their cost estimates so that the cheapest may be chosen. These costs are calculated by the last two modules of the optimizer, the cost model and the estimator-Size allocation.

Statistical Space: This module determines the action execution orders that are to be considered by the Planner for each query sent to it. All such series of actions produce the same query answer, but usually differ in performance. They are usually represented in relational algebra as formulas or in tree form. Because of the algorithmic nature of the objects generated by this module and sent to the Planner, the overall planning stage is characterized as operating at the procedural level.

Structural Space: This module determines the choice of performance that exists for the execution of each set of actions ordered by the field of statistics. This choice is related to the join methods are available for each joint (eg, nested loop, scan and hash them together), as supporting data structures are built on them if / when duplicates are eliminated, and the characteristics of other implementation of this kind, which are determined by the performance of the DBMS. This choice is also linked to evidence any relationship, which is determined by the physical schema of each database stored in its catalog entry Given a Statistical formula or tree from the Statistical Space, this module produces all corresponding complete execution plans, which specify the implementation of each algebraic operator and the use of any indices [6].

Cost Model: This module specify the mathematical formulas that are used to approximate the cost of execution plans. For every different join method, for every different index type access, and in general for every different kind of step that can be found in an execution plan, there is a formula that gives its cost. Given the complexity of many of these steps, most of these formulas are simple approximations of what the system actually does and are based on certain assumptions regarding issues like buffer

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

management, disk-cpu overlap, sequential vs. random I/O, etc. The most important input parameters to a formula are the size of the buffer pool used by the corresponding step, the sizes of relations or indices accessed, and possibly various distributions of values in these relations. While the first one is determined by the DBMS for each query, the other two are estimated by the Size- allocation Estimator.

Size- Allocation Estimator: This module specifies how the sizes (and possibly frequency distributions of attribute values) of database relations and indices as well as (sub) query results are estimated. As mentioned above, these estimates are needed by the Cost Model. The specific estimation approach adopted in this module also determines the form of statistics that need to be maintained in the catalogs of each database, if any [6]

3. Advanced Types of Optimization

In this section, we attempt to provide a concise sight of advanced types of optimization that researchers have proposed over the past few years. The descriptions are based on examples only; further details may be found in the references provided. Furthermore, there are several issues that are not discussed at all due to lack of space, although much interesting work has been done on them, e.g., nested query optimization, rule-based query optimization, query optimizer generators ,object-oriented query optimization, optimization with materialized views, heterogeneous query optimization, recursive query optimization, aggregate query optimization, optimization with expensive selection predicates, and query optimizer validation. Before presenting specific technique consider the following simple relation EMP (empid ,salary, job, department, dno) , DEPT(dno, budget,)

Semantic Query Optimization

Semantic query optimization is a form of optimization mostly related to the Rewriter module. The basic idea lies in using integrity constraints defined in the database to rewrite a given query into semantically equivalent ones [7]. These can then be optimized by the Planner as regular queries and the most efficient plan among all can be used to answer the original query. As a simple example, using a hypothetical SQL-like syntax, consider the following integrity constraint:

assert sal-constraint **on** emp:

salary>200K **where** job = “Assistant professor"

In addition consider the following query:

select empid, subject

from emp, dept

where emp.dno = dept.dno and job = “Assistant professor".

Using the above integrity constraint, the query can be rewritten into a semantically equivalent one to include a selection on sal:

select empid, subject

from emp, dept

where emp.dno = dept.dno and job = “Assistant professor" and salary>200K.

Having the extra selection could help extremely in discovery a fast plan to answer the query if the only index in the database is a B+-tree on emp.sal. On the other hand, it would certainly be a waste if no such index exists. For such reasons, all proposals for semantic query optimization present various heuristics or rules on which rewritings have the potential of being beneficial and should be applied and which not.

Global Query Optimization

So far, we have focused our attention to optimizing individual queries. Quite often, however, multiple queries become available for optimization at the same time, e.g., queries with unions, queries from multiple concurrent users, queries embedded in a single program, or queries in a deductive system. Instead of optimizing each query separately, one may be able to obtain a global plan that, although possibly suboptimal for each individual query, is optimal for the execution of all of them as a group. Several techniques have been proposed for global query optimization [8].

As a simple example of the problem of global optimization consider the following two queries:

select empid, subject

from emp, dept

where emp.dno = dept.dno and job = “Assistant professor ",

select empid

from emp, dept

where emp.dno = dept.dno and budget > 1M

Depending on the sizes of the emp and dept relations and the selectivity’s of the selections, it may well be that computing the entire join once and then applying separately the two selections to obtain the results of the two queries is more efficient than doing the join twice, each time taking into account the corresponding selection. Developing Planner modules that would examine all the available global plans and identify the optimal one is the goal of global/multiple query optimizers.

Parametric Query Optimization

As mentioned earlier, embedded queries are typically optimized once at compile time and are executed multiple times at run time. Because of this temporal separation between optimization and execution, the values of various parameters that are used during optimization may be very different during execution. This may make the chosen plan invalid (e.g., if indices used in the plan are no longer

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

available) or simply not optimal (e.g., if the number of available buffer pages or operator selectivity's have changed, or if new indices have become available). To address this issue, 31several techniques [9,10,11] have been proposed that use various search strategies (e.g., randomized algorithms [10] or the strategy of Volcano [11]) to optimize queries as much as possible at compile time taking into account all possible values that interesting parameters may have at run time. These techniques use the actual parameter values at run time, and simply pick the plan that was found optimal for them with little or no overhead. Of a drastically different flavor is the technique of Rdb/VMS [12], where by dynamically monitoring how the probability distribution of plan costs changes, plan switching may actually occur during query execution**.**

Conclusion

To a large extent, the success of a DBMS lies in the quality, functionality, and sophistication of its query optimizer, since that determines much of the system's performance. In this paper, we have given a bird's eye view of query optimization. We have presented an abstraction of the architecture of a query optimizer and focused on the techniques currently used by most commercial systems for its various modules. In addition, we have provided a glimpse of advanced issues in query optimization, whose solutions have not yet found their way into practical systems, but could certainly do so in the future.

References

[1] J. Gray, D.T. Liu, M.A. Nieto-Santisteban, A. Szalay, D.J. DeWitt, and G. Heber, "Scientific data management in the coming decade", SIGMOD Record 34(4), pp. 34-41, 2005.

[2] Ruslan Fomkin and Tore Risch 1997 "Cost-based Optimization of Complex Scientific Queries", Department of Information Technology, Uppsala University

[3] C. Hansen, N. Gollub, K.Assamagan, and T. Ekelöf, "Discovery potential for a charged Higgs boson decaying in the chargino-neutralino channel of the ATLAS detector at the LHC", Eur.Phys.J. C44S2, pp. 1-9, 2005.

[4] Melton, J., Simon A. Understanding The New SQL: A Complete

[5] Graefe G. Query Evaluation Techniques for Large Databases. In ACM Computing Surveys: Vol 25, No 2., June 1993.

[6] Yannis E. Ioannidis," Query optimization" Computer Sciences Department,University of Wisconsin Madison, WI 53706

[7] J. J. King. Quits: A system for semantic query optimization in relational databases. In Proc. of the 7th Int. VLDB Conference , pages 510{517, Cannes, France, August 1981.

[8] T. Cells. Multiple query optimization. ACM-TODS, 13(1):23{52, March 1988.

[9] G. Graefe and K. Ward. Dynamic query evaluation plans. In Proc. ACM-SIGMOD Conference on the Management of Data, pages 358-366, Portland, OR, May 1989.

[10] Y. Ioannidis, RNg, K. Shim, and T. K. Sellis. Parametric query optimization. In Proc. 18th Int. VLDB Conference, pages 103{114, Vancouver, BC, August 1992.

[11] R. Cole and G. Graefe. Optimization of dynamic query evaluation plans. In Proc .ACM-SIGMOD Conference on the Management of Data, pages 150{160, Minneapolis,MN, June 1994.

[12] G. Antoshenkov. Dynamic query optimization in Rdb/VMS. In Proc. IEEE Int. Coference on Data Engineering, pages 538{547, Vienna, Austria, March 1993.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A New Improved Algorithm for Distributed Databases

K.Karpagam
Assistant Professor, Dept of Computer Science,
H.H. The Rajah's College (Autonomous),
(Affiliated to Bharathidasan University, Tiruchirappalli)
Pudukkottai, Tamil Nadu, India.

Dr.R.Balasubramanian
Dean, Faculty of Computer Applications,
EBET Knowledge Park,
Tirupur, Tamil Nadu, India.

***Abstract*—The development of web, data stores from disparate sources has contributed to the growth of very large data sources and distributed systems. Large amounts of data are stored in distributed databases, since it is difficult to store these data in single place on account of communication, efficiency and security. Researches on mining association rules in distributed databases have more relevance in today's world. Recently, as the need to mine patterns across distributed databases has grown, Distributed Association Rule Mining algorithms have gained importance. Research was conducted on mining association rules in the distributed database system and classical Apriori algorithm was extended based on transactional database system. The Association Rule mining and extraction of data in distributed sources combined with the obstacles involved in creating and maintaining central repositories motivates the need for effective distributed information extraction and mining techniques. We present a new distributed association rule mining algorithm for distributed databases (NIADD). Theoretical analysis reveals a minimal error probability than a sequential algorithm. Unlike existing algorithms, NIADD requires neither knowledge of a global schema nor that the distribution of data in the databases.**

Keywords- Distributed Data Mining, Distributed Association Rules

I. INTRODUCTION

The essence of KDD is Acquisition of knowledge. Organizations have a need for data mining, since Data mining is the process of non-trivial extraction of implicit, previously unknown and potentially useful information from historical data. Mining association rules is one of the most important aspects in data mining. Association rules Mining (ARM) can predict occurrences of related. Many applications use Data Mining for rankings of products or data based decisions. The main task of every ARM algorithm is to discover the sets of items that frequently appear together (Frequent item sets). Many organizations are geographically distributed and merging data from locations into a centralized site has its own cost and time implications.

Parallel processing is important in the world of database computing. Databases often grow to enormous sizes and are accessed by more and more users. This volume strains the ability of single-processors systems. Many organizations are turning to parallel processing technologies for performance, scalability, and reliability. Much progress has also been made in parallelized algorithms. The algorithms have been effective in reducing the number of database scans required for the task. Many algorithms were proposed which take advantage of the speed in network or the memory or parallel computers. Parallel computers are costly. The alternative is distributed algorithms, which can run on lesser costing clusters of PCs. Algorithms suitable for such systems include the CD and FDM algorithms [2, 3], both parallelized versions of Apriori. CD and FDM algorithms did not scale well on the increase of the clustered PC's [4].

II. DISTRIBUTED DATABASES

There are many reasons for organizations to implement a Distributed Database system. A distributed database (DDB) is a collection of multiple, logically interrelated databases distributed over a computer network. The distribution of databases on a network achieves the advantages of performance, reliability, availability and modularity that are inherent in distributed systems. Many organizations which use relational database management system (RDBMS) have multiple databases. Organizations have their own reasons for using more than a single database in a distributed architecture as in Figure 1. Distributed databases are used in scenarios where each database is associated with particular business functions like manufacturing. Databases may also be implemented based on geographical boundaries like headquarters and branch offices.

The users accessing these databases access the same data in different ways. The relationship between multiple databases is part of a well-planned architecture, in which distributed databases are designed and implemented. A distributed database system helps organizations serve their objectives like *Availability, Data collection, extraction* and *Maintenance*. Oracle an RDBMS has inter database connectivity with SQL*Net. Oracle also supports Distributed Databases by *Advanced replication* or *multi-master replication.* Advanced replication is used to deliver high availability. Advanced replication involves numerous databases. Oracle's parallel query option (PQO) is a technology that divides complicated or long-running queries into many small queries which are executed independently.

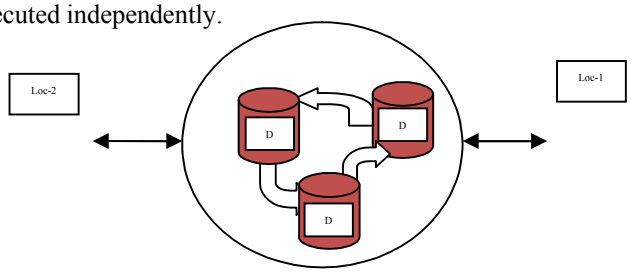


Figure 1 Distributed Database system

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. BENEFITS OF DISTRIBUTED DATABASES

The separation of the various system components, especially the separation of application servers from database servers, yields tremendous benefits in terms of cost, management, and performance. A machine's optimal configuration is a function of its workload. Machines that house web servers, for example, need to service a high volume of small transactions, whereas a database server with a data warehouse has to service a relatively low volume of large transactions (i.e., complex queries). A distributed architecture is less drastic than an environment in which databases and applications are maintained on the same machine. Location transparency implies neither applications nor users need to be concerned with the logistics of where data actually resides. Distributed databases allow various locations to share their data. The components of the distributed architecture are completely independent of one another, which mean that every site can be maintained independently. Oracle Database's Database links makes Distributed Databases to be linked together.

For Example

CREATE PUBLIC DATABASE LINK LOC1.ORG.COM USING hq.ORG.COM.

An example of a Distributed query would be

SELECT emplyeename, Department
from EmployeeTable E, DepartmentTable@hq.ORG.COM D
WHERE E.empno = D.empno

IV. PROBLEM DEFINITION

Association Rule mining is an important data mining tool used in many applications. Association rule mining finds interesting associations and/or correlation relationships among large sets of data. Association rules show attributes value conditions that occur frequently together in a given dataset. A typical and widely-used example of association rule mining is market basket analysis. For example, data collected in supermarkets having large number of transactions. Answering a question like set of items purchased often is not so easy. Association rules provide information of this type in the form of "if-then" statements. The rules computed from the data are based on probability. Association rules are one of the most common techniques of data mining for local-pattern discovery in unsupervised learning systems [5]. A random sample of the database is used to predict all the frequent item sets, which are then validated in a single database scan. Because this approach is probabilistic not only the frequent item sets are counted in the scan but also the negative border (an itemset is in the negative border if it is not frequent but all its "neighbors" in the candidate itemset are frequent) is considered. When the scan reveals item sets in the negative border are frequent, a second scan is performed to discover whether any superset of these item sets is also frequent. The number of scans increases the time complexity and more so in Distributed Databases. The purpose of this paper is to introduce a new Mining Algorithm for Distributed Databases. A large number of parameters affect the performance of distributed queries.Relations involved in a distributed query may be fragmented and/or replicated. With many sites to access, query response time may become very high.

V. PREVIOUS WORK

Researchers and practitioners have been interested in distributed database systems since 1970s. At that time, the main focus was on supporting distributed data management for large corporations and organizations that kept their data at different locations. Distributed data processing is both feasible and needed. Almost all major database system vendors offer products to support distributed data processing (e.g.,IBM, Informix, Microsoft, Oracle, Sybase). Since its introduction in 1993 [5], the ARM problem has been studied intensively. Many algorithms, representing several different approaches, were suggested. Some algorithms, such as Apriori, Partition, DHP, DIC, and FP-growth [6, 7, 8, 9, 10], are bottom-up, starting from item sets of size and working up. Others, like Pincer-Search [11], use a hybrid approach, trying to guess large item sets at an early stage. Most algorithms, including those cited above, adhere to the original problem definition, while others search for different kinds of rules [9, 12, 13]. Algorithms for the Distributed ARM can be viewed as parallelizations of sequential ARM algorithms. The CD, FDM, and DDM [2, 3, 14] algorithms parallelize Apriori [6], and PDM [15] parallelizes DHP [16]. The parallel algorithms use the architecture of the parallel machine, where shared memory is used [17].

VI. APRIORI ALGORITHM FOR FINDING FREQUENT ITEM SETS

The Apriori algorithm for finding frequent item sets and is explained. Let k-item set be an item set which consists of k items, then Frequent itemset F_k is an itemset with sufficient support and a large itemset is denoted by L_k. Let c_k be a set of candidate k-item sets. The Apriori property is, if an item X is joined with item Y, then

Support(X U Y) = min(Support(X), Support(Y))

The first iteration is to find L1, all single items with Support > threshold. The second iteration would be to find L2 using L1. The iterations would continue until no more frequent k item sets can be found. Each iteration i consist of two phases:

Candidate generation - Construct a candidate set of large item sets

Counting and selection - Count the number of occurrences of each candidate item set and Determine large item sets based on predetermined support

Set L_k is defined as the set containing the frequent k item sets which satisfy

Support > threshold.

$L_k * L_k$ is defined as:

$L_k * L_k = \{X \cup Y$, where X, Y belong to L_k and $|X \cap Y| = k-1\}$.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

VII. DISTRIBUTED ALGORITHMS IN ASSOCIATION RULES

A. PARALLEL PROCESSING FOR DATABASES

Three issues drive the use of parallel processing in database environments namely speed of performance, scalability and availability. Increase in Database size increases the complexity of queries. Organizations need to effectively scale their systems to match the Database growth. With the increasing use of the Internet, companies need to accommodate users 24 hours a day. Most parallel or distributed association rule algorithms parallelize either the data or the candidates. Other dimensions in differentiating the parallel association rule algorithms are the load-balancing approach used and the architecture. The data parallelism algorithms require that memory at each processor be large enough to store all candidates at each scan. The task parallel algorithms adapt to the amount of available memory at each site, since all partitions of the candidates may not be of the same size. The only restriction is that the total size of all candidates be small enough to fit into the total size of memory in all processors combined.

B. *FDM ALGORITHM*

The FDM (Fast Distributed Algorithm for Data Mining) algorithm, proposed in (Cheung *et al.* 1996) has the following distinguishing characteristics:

Candidate set generation is Apriori-like.

After the candidate sets are generated, different types of reduction techniques are applied, namely a local reduction and a global reduction, to eliminate some candidates in each site.

The FDM algorithm is shown below.

Input:
DBi //database partition at each site *Si*
Output:
L //set of all globally large itemsets
Algorithm:
Iteratively execute the following program fragment (for the *k*th iteration) distributively at each site *Si*.

The algorithm terminates when either $L(k) = \emptyset$, or the set of candidate sets

$CG(k) = \emptyset$.

if $k = 1$ **then**

$Ti(1) = get\_local\_count(DBi, \emptyset, 1)$

else {

$CG(k) = \cup n$

$i=1\ CGi(k) = \cup n$

$i=1\ Apriori\_gen(GLi(k-1))$
$Ti(k) = get\_local\_count(DBi, CG(k), i)$ }

for each $X \in Ti(k)$ **do**

if $X.supi \geq s \times Di$ **then**
for $j = 1$ to n **do**
if $polling\_site(X) = Sj$ **then**
insert $\langle X, X.supi \rangle$ into $LLi,j(k)$
for $j = 1$ to n **do**
send $LLi,j(k)$ to site Sj
for $j = 1$ to n **do** {
receive $LLj,i(k)$
for each $X \in LLj,i(k)$ **do** {

if $X \notin LPi(k)$ **then**
insert X into $LPi(k)$
update $X.large\_sites$ } }
for each $X \in LPi(k)$ **do**
send_polling_request(X);
reply_polling_request(Ti(k))
for each $X \in LPi(k)$ **do** {
receive $X.supj$ from sites Sj
where $Sj \notin X.large\_sites$

$X.sup = \Sigma n$
$i=1\ X.supi$
if $X.sup \geq s \times D$ **then**
insert X into $Gi(k)$ }

1. broadcast $Gi(k)$

receive $Gj(k)$ from all other sites Sj, $(j \neq i)$

$L(k) = \cup n$
$i=1\ Gi(k)$
divide $L(k)$ into $GLi(k)$, $(I = 1,\ldots,n)$

1. **return** $L(k)$.

VIII. NIADD ALGORITHM

Parallel processing involves taking a large task, dividing it into several smaller tasks, and then working on each of those smaller tasks simultaneously. The goal of this divide-and-conquer approach is to complete the larger task in less time than it would have taken to do it in one large chunk. In parallel computing, Computer hardware is designed to work with multiple processors and provides a means of communication between those processors. Application software has to break large tasks into multiple smaller tasks and perform in parallel. NIADD is algorithm striving to get the maximum advantage of using the RDBMS like parallel processing.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. NIADD CHARECTERISTICS

The NIADD (New Improved Algorithm for Distributed Databases) algorithm has the following distinguishing characteristics. Candidate set generation is Apriori-like, but frequent item sets generated with Minimum support reduces the set of candidates commonly. The Algorithm uses the power of Oracle and its Memory Architectures to attain speed. An oracle query is executed with the support% as a parameter for reduction of candidates.

B. NIADD ALORITHM

Let D be a transactional database with T transactions at Locations L1, L2, ..., Ln. The databases are { D_1, D_2, …. D_i }. Let T_1, T_2, …. T_i be the Transactions at each Location. Let F_k be the set of Common Frequent item sets. Let Min Support be Defined as a percentage and the Criteria to Filter Transactions where $T_{1..n} \geq$ Min Support. The main goal of a distributed association rules mining algorithm is finding the globally frequent item sets F. The NIADD Algorithm is defined as

```
for each D1..n do //where 1..n = Di
        for each T1..n □ Di do
            if Ti(support) ≥ Min Support then
                    Select Ti into Fk
             end if
        end for
    end for
```

IX. CHALLENGES

Mining Distributed Databases has to address the problem of large-scale data mining. It has to speed up and scale up data mining algorithms.

Challenges:

- Multiple scans of transaction database
- Huge number of candidates
- Tedious workload of support counting for candidates

Possible Solutions:

- Reduce passes of transaction database scans
- Shrink number of candidates
- Facilitate support counting of candidates

The itemsets can be reduced by reducing the number of transactions to be scanned by Transaction reduction. Any transaction which does not contain frequent k-itemsets cannot contain any frequent (k + 1) - itemsets. The transaction can be filtered from further scans. Partitioning techniques which require two database scans to mine the frequent itemsets can be used. The First Phase subdivides the transactions of D into n non-overlapping partitions. If the minimum support threshold for transactions in D is min sup, then the minimum itemset support count for a partition is min sup x the number of transactions in that partition. For each partition, all frequent itemsets within the partition are found. These are referred to as local frequent itemsets. The procedure employs a special data structure which, for each itemset, records the TID's of the transactions containing the items in the itemset. This allows it to find all of the local frequent k-itemsets, for k = 1:2, in just one scan of the database. In the second Phase, a second scan of D is conducted in which the actual support of each candidate is assessed in order to determine the global frequent itemsets.

X. PERFORMANCE AND RESULTS

NIADD Finds sequences of transactions associated over a support factor. The goal of pattern analysis is to find sequences of itemsets. A transaction sequence can contain an itemset sequence if each itemset is contained in one transaction. i.e. If the ith itemset in the itemset sequence is contained in transaction j in the transaction sequence, then the (i + 1)th itemset in the itemset sequence is contained in a transaction numbered greater than j. The support of an itemset sequence is the percentage of transaction sequences that contain it. The data set used for testing the performance the NIADD algorithm was generated by setting the maximum number locations as Three. The algorithms were implemented in Oracle 10g and the support factor was varied between 0.5% and 5%. Figure 1 shows the performance of the algorithms depending on the number Transactions and Distributed Databases count. To decrease the execution time, filters (Min Support Percentage) were increased. It was found there was a noticeable improvement in the performance of the algorithms with increments in the support factor.

```
SELECT
 EmpId , EmpName, EmpBasic
  FROM emp@loc1.db
  Union
EmpId , EmpName, EmpBasic
  FROM emp@loc2.db
  Union
EmpId , EmpName, EmpBasic
            FROM emp@loc3.db
  Where EmpBasic > 3000
```

A. ANALYSIS AND OBSERVATIONS

1. The time taken to retrieve a row from a Very Large Database is less than 1 second.
2. The time taken increases with the number of rows
3. The time taken on multiple item attributes is unimaginable.
4. The information retrieval is directly proportional to the number of Transactions in the database.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. SOLUTION

Goal is to identify Frequent Item sets in Distributed Databases

1. Determining What to Select
   - The Attributes of an Item is translated to Columns of the Transactions.
2. Selecting frequent Item sets.

C. EXPERIMENTAL REULTS OF NIADD

Experiments were conducted to compare response times obtained with FDM and NIADD on the Distributed Databases. It was noticed; increase in the Min Support decreased the computation time.

TABLE 1: FREQUENT ITEMSET RETRIEVAL TIME OF FDM AND NIADD BASED DISTRIBUTED DATABASES

SL.No.	No. of Databases	FDM in Secs	NIADD in Secs
2	1	7.6	8.92
3	2	12.1	13.6
4	3	16.2	17.6

TABLE 2: FREQUENT ITEMSET RETRIEVAL TIME OF FDM AND NIADD BASED SUPPORT FACTOR

SL.No.	Support %	FDM in Secs	NIADD in Secs
1	0.5	7.6	8.92
2	1	3.838	4.46892
3	2	0.97869	1.1217
4	3	0.16800845	0.18807
5	5	0.01764089	0.019

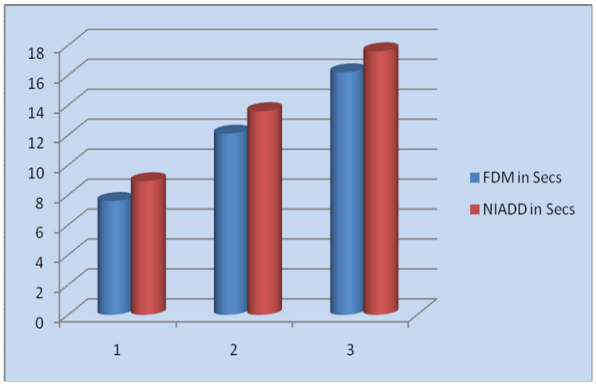


Figure 2 - Response times obtained with FDM and NIADD based on Number of Databases

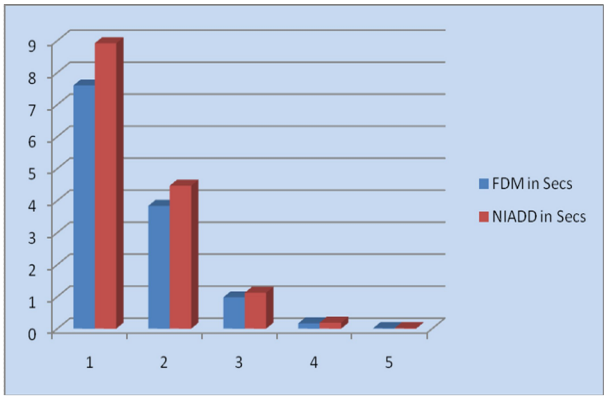


Figure 3 - Response times obtained with FDM and NIADD based on Min Support %

The data set used for testing the performance of the two algorithms, NIADD and FDM, was generated according to (Agrawal and Shrikant 1994), by setting the number of items N = 100, and the increasing the support factor. To test the described algorithms, 1 to 3 Databases were used. The algorithms were implemented in Oracle 10g. To study the algorithms the support factor was varied between 0.5% and 5%. A first result, obtained by testing the two algorithms on data sets with 1000 to 5000 transactions and, as mentioned before, using between 1 and 3 Databases with a support factor of a maximum of 5%. The performance of the algorithm depends on the support factor % and the number of transactions. For a data set with 4500 transactions that was distributed on three Databases, an execution time of just 8.92 seconds for the NIADD algorithm and 7.6 seconds for the FDM algorithm. The data set with 1000 transactions was distributed on 2 sites the execution time for the NIADD algorithm was 68 second and for the FDM algorithm 60 seconds, while the same data set distributed on 3 sites the execution time has raised to 88 seconds for the NIADD algorithm and to 80 seconds for the FDM algorithm. The FDM performance increased since it used the respective processors at locations of the databases. It is noticeable that the performance of the algorithms increases with the support factor, but the FDM algorithm presents a better performance than the NIADD algorithm. From the experiments made, resulted a good scalability for the NIADD and FDM algorithms, relative to different support factors for a large data set. The distributed mining algorithms can be used on distributed databases, as well as for mining large databases by partitioning them between sites and processing them in a distributed manner. The high flexibility, the scalability, the small cost/performance ratio and the connectivity of a distributed system make them an ideal platform for data mining.

XI. CONCLUSION

Finding all frequent item sets in a database in real-world applications, is a problem since the transactions in the database can be very large scaling up to 10 terabytes of data. Frequent item sets increases exponentially based on the number of different items. Experimental results show, mining algorithms

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

do not perform evenly when implemented in Oracle, demarcating space for performance improvements. The algorithms determine all candidates in Distributed Database architecture. For any frequent item in an item set, candidates that are immediate supersets of the item need to be determined. In this paper a new improved algorithm, NIADD is presented. The new algorithm is compared with FDM. The results indicate that the NIADD algorithm is well suited and effective for finding frequent item sets with less execution time. Also, increasing the support factor proportionately increases the performance of the algorithm. These results show the fact that the increase in Min Support is done relative to the Transaction values in the Database's dataset. The NIADD can be used on distributed databases, as well as for mining large volumes of data based on the Memory of the main site. This leaves scope for improvement of the NIADD by using multiple-processor's memory like the FDM.

REFERENCES

[1] Lan H. Witten, Eibe Frank, "Data Mining",China Machine Press, Beijing, 2003.

[2] R. Agrawal and J. Shafer, "Parallel mining of association rules. IEEE Transactions on Knowledge and Data Engineering", pages 962 . 969, 1996.

[3] D. Cheung, J. Han, V. Ng, A. Fu, and Y. Fu., "A fast distributed algorithm for mining association rules", In Proc. Of 1996 Int'l. Conf. on Parallel and Distributed Information Systems, pages 31 . 44, Miami Beach, Florida, December 1996

[4] A. Schuster and R. Wolff, "Communication-efficient distributed mining of association rules", In Proc. of the 2001 ACM SIGMOD Int'l. Conference on Management of Data, pages 473 . 484, Santa Barbara, California, May 2001.

[5] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases", In Proc. of the 1993 ACM SIGMOD Int'l. Conference on Management of Data, pages 207.216, Washington, D.C., June 1993.

[6] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In Proc. of the 20th Int'l. Conference on Very Large Databases (VLDB'94), pages 487 . 499, Santiago, Chile, September 1994.

[7] A. Savasere, E. Omiecinski, and S. B. Navathe, "An efficient algorithm for mining association rules in large databases", The VLDB Journal, pages 432.444, 1995.

[8] J. S. Park, M.-S. Chen, and P. S. Yu, "An effective hashbased algorithm for mining association rules", In Proc. Of ACM SIGMOD Int'l. Conference on Management of Data, pages 175 . 186, San Jose, California, May 1995.

[9] S. Brin, R. Motwani, J. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data", SIGMOD Record, 6(2):255.264, June 1997.

[10] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation", Technical Report 99-12, Simon Fraser University, October 1999.

[11] D.I. Lin and Z. M. Kedem. Pincer search, "A new algorithm for discovering the maximum frequent set", In Extending Database Technology, pages 105.119, 1998.

[12] R. Srikant and R. Agrawal, "Mining generalized association rules", In Proc. of the 20th Int'l. Conference on Very Large Databases (VLDB'94), pages 407 . 419, Santiago, Chile, September 1994.

[13] J. Pei and J. Han, "Can we push more constraints into frequent pattern mining?", In Proc. of the ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, pages 350.354, Boston, MA, 2000.

[14] A. Schuster and R. Wolff, "Communication-ef_cient distributed mining of association rules", In Proc. of the 2001 ACM SIGMOD Int'l. Conference on Management of Data, pages 473 . 484, Santa Barbara, California, May 2001.

[15] J. S. Park, M.-S. Chen, and P. S. Yu., "Efficient parallel data mining for association rules", In Proc. of ACM Int'l. Conference on Information and Knowledge Management, pages 31.36, Baltimore, MD, November 1995.

[16] J. S. Park, M.-S. Chen, and P. S. Yu, "An effective hash based algorithm for mining association rules", In Proc. Of ACM SIGMOD Int'l. Conference on Management of Data, pages 175 . 186, San Jose, California, May 1995.

[17] O. R. Zaiane, M. El-Hajj, and P. Lu, "Fast parallel association rules mining without candidacy generation", In IEEE 2001 International Conference on Data Mining(ICDM'2001), pages 665.668, 2001.

AUTHORS PROFILE

K.Karpagam, *M.Sc., M.Phil.,* Assistant Professor, Dept of Computer Science, H.H. The Rajah's College(Autonomous), Pudukkottai, Tamil Nadu, India (affiliated to Bharathidasan University, Tiruchirappalli). She has to her credit 13 years of teaching experience and currently pursuing Ph.D. research at Mother Teresa University, Kodaikanal. Tamil Nadu, India.

email:- kkarpaga05@gmail.com

Dr.R.Balasubramanian, Ph.D., M.Phil(Maths)., M.Phil.(CS)., M.Phil(Mgt.)., M.S., MBA., M.Sc., MADE., PGDCA., PGDIM., PGDOM., PGDOM., PGDHE., DE., DIM., CCP., MADE., PGDCA., PGDIM., PGDOM., PGDOM., PGDHE., DE., DIM., CCP., Dean, Faculty of Computer Applications, EBET, Nathakadaiyur, Tirupur, Tamil Nadu. , has more than 34 years of teaching experience in Tamil Nadu Government Collegiate Educational Service at various capacities as SG Lecturer in Maths (24 years), SG Lecturer and HoD of Computer Science (9 years) and Principal (1 year). He was formerly serving as Principal at Raja Doraisingam Government Arts College, Sivagangai. He had been Chairman of PG Board of Computer Science of Bharathidasan University, Trichy for a period of 3 years.

He is a recognized guide in Computer Science, Mathematics and Education. He has wide research experience in areas like Computer Science, Education, Mathematics and Management Science. He has produced 3 doctorates in computer Science and he is presently guiding 15 Ph.D Scholars in Computer Science and 2 Ph.D Scholars in Education of various universities. He has completed many projects including two major projects of UGC.

Data mining applications in modeling Transshipment delays of Cargo ships

P. Oliver jayaprakash
Ph.D student, Division of Transportation engg,
Dept.of Civil engineering, Anna University,
Chennai, Tamilnadu, India
e-mail:sendmailtooliver@yahoo.com

Dr.K.Gunasekaran,
Associate Professor, Division of Transportation engg.
Dept.of Civil engineering, Anna University,
Chennai, Tamilnadu, India
e-mail:kgunasekaran@hotmail.com

Dr.S.Muralidharan
Professor, Dept.of EEE
Mepco schlenk engineering college,
Sivakasi, Tamilnadu, India.
e-mail:yes_murali@yahoo.com

***Abstract*—The Data mining methods have a plenty of applications in various fields of engineering. The present application area is the Port operations and management. Conventionally port performance was assessed by the ship turnaround time, a marker of cargo handling efficiency. It is a time used up at port for transshipment of cargo and servicing. During the transshipment and servicing, delays were inevitable and occur predominantly; The major delay happening at port was due to the non-availability of trucks for evacuation of cargo from port wharf to the warehouses. Hence, modeling the delay occurrences in port operations had to be done, so as to control the ship's turnaround time at the port to prevent additional demurrage charges. The objective of this paper was to study the variety of delays caused during the port processes and to model it using Data mining techniques.**

Keywordst; Data mining techniques, Transshipment delays, Shunt trucks, Artificial neural network, Nonlinear analysis.

I. INTRODUCTION

The growing volume of Port related transhipment data raises many challenges, one is to extract, store, organize, and use the relevant knowledge generated from those data sets. The data content with differing time periods could be deployed for various engineering applications. The innovations that occur in computing infrastructure and the emergence of data mining tools have an impact on decision making related port shipment operations. The growing demand for data mining has led to the development of many algorithms that extract knowledge and features such as missing data values, correlation, trend and pattern, etc. from a large scale databases. Data mining techniques play a crucial role in several fields of engineering applications. They help the managers in formatting the data collected over an issue and collecting the potential information out of the data through preprocessing and warehousing tools. The conventional MLR models were replaced by Nonlinear and ANN models to do the prediction of future variable values related to the complex systems, even with the minimum data because of their accuracy and reliability in results. This paper focus on the application of data mining techniques in processing the Non-containerized ships related transhipment delays and model it using various models such as MLR, NLR and ANN. A ship's service time, which affects quantum of the consignments imported and exported in a particular time period, was much influenced by berth planning and allocation. Also, it affects the Ship turnaround time, since the vessels' length of stay at port was decided by it. The delay caused by shunt trucks at port gates was one of the crucial issues faced by the Port authorities. The cargo evacuation period was influenced by shunt trucks turnaround time. The turnaround time of a truck was estimated as the time taken to evacuate the cargo completely from the port's quay or wharf to the company warehouses located in the port outer area. Port terminals trying to minimise the truck turnaround time, so as to reduce the inland transportation cost of cargo evacuation. The delay component was significant, varying and high in developing countries compared to the efficient ports of developed countries.

The export or import of commodity was done by the procedures of port system given in the Figure 1. The major factors affecting the ship servicing delay were lengthy port operational procedures in importing or exporting the cargo, ship related delays (not related to port) and port related delays and delays due to carriers. Hence, it was necessary to analyse the causes behind delays and to formulate strategies to minimise it.

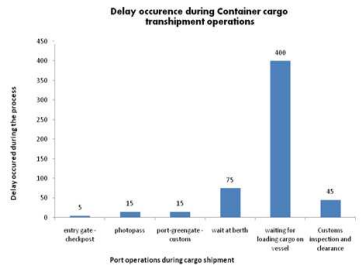


Figure 1 Operations in Non-containerised cargo

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The list of procedures related to truck shunt operations to evacuate the cargo is given below;

Procedures involved in transshipment operations

- Prepare transit clearance
- Inland transportation
- Transport waiting for pickup &loading
- Wait at port entry
- Wait at berth
- Terminal handling activities

II. PAST LITERATURE

Ravikumar [1] compared the various data masking techniques such as encryption, shuffling, scrubling, etc and its wide applications in various industries to secure data from hacking and discussed the advantages of Random Replacement as one of the standard method for data masking with the highest order of security. Mohammad behrouzian [2] discussed the advantages, limitations and applications of data mining in various industries and the banking industry, especially in the customer relationship management. According to Krishnamurthy [3] data mining is an interface among the broad disciplines like statistics, computer science and artificial intelligence, machine learning and data base management,etc., Kusiak [4] introduced the concepts of machine learning and data mining and presented the case studies of its applications in industrial, medical, and pharmaceutical domains.

Chang Qian Gua [5] discussed the gate capacity of container terminals and built a multiserver queuing model to quantify and optimize the truck delays. Wenjuan Zhao and Anne V. Good child [6] quantified the benefits of truck information that can significantly improve crane productivity and reduce truck delay for those terminals operating with intensive container stacking. Unctad report [7] suggests various port efficiency parameters to rank the berth productivity. The parameters used were, average ship berth output, delays at berth, duration of waiting for berth and turn-round time. Nathan Huynh [8] developed a methodology for examining the sources of delay of dray trucks at container terminals and offered specialized solutions using decision trees, a data mining technique. U. Bugaric [9] developed a simulation model to optimize the capacity of the Bulk cargo river terminals by reducing transshipment delay, without investing on capital costs. Mohammed ali [10] simulated the critical conditions, when ships were delayed at offshore and containers were shifted to port by barges; Kasypi mokhtar [11] built a regression model for vessel turnaround time considering the Transshipment delays and number of gangs employed per shift, etc. Simeon Djankov [12] segregated the pre-shipment activities such as inspection and technical clearance; inland carriage and handling; terminal handling, including storage, Customs and technical control. And, he conducted an opinion survey to estimate the delay caused in document clearance, fees payment and approval processes. F. Soriguera, D. Espinet, F. Robuste [13] optimized the internal transport cycle using an algorithm, by investigating the sub systems such as landside transport, storage of containers in a marine container terminal. Brian M. Lewis, Alan L. Erera, and Chelsea C. White [14] designed a markov process based decision model to help stakeholders quantify the productivity impacts of temporary closures of a terminal. He demonstrated the uses of decision trees to gain insight into their operations instead of exhaustive data analysis. Rajeev namboothiri [15] studied the fleet operations management of drayage trucks in a port. Truck congestion at ports may lead to serious inefficiencies in drayage operations. H.Murat Celik [16] developed three different ANN models for freight distribution of short term inter-regional commodity flows among 48 continental states of US, utilizing 1993 commodity survey data. Peter B. Marlow [17] proposed a new concept of agile ports, to measure the port performance by including quantitative and qualitative parameters. Rahim F. Benekohal, Yoassry M. El-Zohairy, and Stanley Wang [18] evaluated the effectiveness of an automated bypass system in minimizing the traffic congestion with the use of automatic vehicle identification and Low speed weight in motion around a weigh station in Illinois to facilitate preclearance for trucks at the weigh station. Jose L. Tongzon [19] built a port performance model to predict efficiency of transshipment operations. This present research focus on Bulk ports handling Non-containerized cargo ships. The transshipment delay data was used for building a predictive model for the future ship delays.

TABLE I
SUMMARY OF TRANSHIPMENT DELAY DATA

Variable	Mean	S.D	Min.	Max.
X_1	102	55	34	504
X_2	0.88	0.36	0.26	1.74
X_3	0.03	0.04	0.00	0.08
X_4	0.28	0.12	0.05	0.72
X_5	27.00	25.00	5.00	80.00
X_6	2.35	1.44	0.33	5.78
X_7	0.04	0.03	0.01	0.18
X_8	0.038	0.026	0.01	0.18
Y	0.18	0.09	0.00	0.35

Where,
Y = Transshipment delay of Non-containerized cargo.
X_1=Number of evacuation trucks,X_2=Truck travel time,X_3=Gang nonworking time,X_4=Truck shunting duration,X_5=Trip distance ,X_6=Berth Time at berths,X_7=Waiting time at berth,X_8= other miscellaneous delays .

III. DATA COLLECTION & ANALYSIS

The noncontainerised cargo ship data were collected for the past five years from 2004 to 2009 from various sources including India seaports [20, 21&22] for a study port. The data comprised of number of ship cranes, number of trucks required to evacuate, crane productivity, truck travel time, idle time, gang idle time, truck shunt time, truck trip distance,

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

delay caused at berth and the gross delay, ship waiting time for berth outside the channel, time spent on berth (berthing time) and ship turnaround time. The summary of ship delay data and the methodology of the study were presented in Table 1 & Figure 2.

A. preprocessing,Correlation and Trend

The collected data was preprocessed using data transformation algorithm and the missing values in the database were filled and the descriptive statistics was estimated. The average crane working time was 5.93 hours per day and mean gang idle time was 0.03 days. The mean berthing time was 2.3 days and the mean ship turnaround time was 2.71 days. A multivariate analysis was done to estimate the correlation among dependent and independent variables. The correlation matrix showing the correlation among the variables was presented in Table II. The average Crane efficiency at the study port was 19616 Tonnes per day; average ship waiting time at berth was 0.04 day and the mean crane productivity was 7.67 Tonnes per hour. The average number of trucks required for evacuation was 104; the mean truck travel time was 0.88 hour mean delay caused to the ship at the port was 0.18 day.

To study the relationship between the independent variables and dependant variable, correlation analysis was carried out and the results were presented in Table II. The independent variable, transshipment delay is highly correlated with Delay caused at storage area and by gang /workforce and further it was correlated with the ship berthing time at port. Also, it was significantly correlated to the number of evacuation trucks, travel time of truck and trip distance, etc.

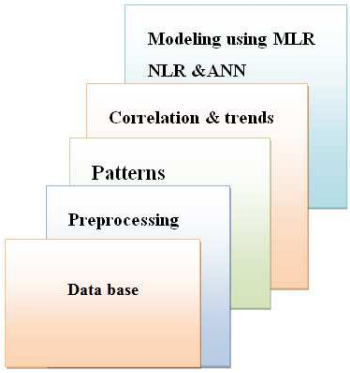


Figure 2 Methodology of the study

IV. DATA COLLECTION & ANALYSIS

Using the historical data on Transhipment delay collected, an ANN model was built, to study the relationship between Transhipment delay and other influencing parameters. Also, a MLR model and a multivariate nonlinear regression model were built for the above data and statistical performance
and prediction accuracy of models were compared and the outcomes were presented.

A *Artificial neural network modeling*

An artificial neural network was an emulation of biological neural system which could learn and calibrate itself. It was developed with a systematic step-by-step procedure to optimize a criterion, the learning rule. The input data and output training was fundamental for these networks to get an optimized output. The neural network was good at studying patterns among the input data and learns. The prediction accuracy increases with the number of learning cycles and iterations. The estimation of Gross transhipment delay caused to the commodity ship tends to vary with type of cargo, season, shipment size and other miscellaneous factors, the most popular and accurate prediction technique; MATLAB's Back propagation neural network (BPNN) module was utilized to predict the Transhipment delay faced by non-containerised ships from the past data. Figure 3 present the hidden layer and architecture of BPNN. The ANN based model was built and training was done using three years' of past data and for testing & production, the two years data were used. The inputs, fleet strength of evacuation trucks, truck travel time, delay due to gang -workforce, idle time, shunting time, trip distance, berth time, delay at storage area were given as batch files and the script programming was used to run neural network model with adequate hidden neurons and the

TABLE II
CORRELATION VALUES BETWEEN VARIABLES

	X_1	X_2	X_3	X_4	X_5
X_1	1.00	-0.98	-0.35	-0.50	-0.18
X_2	-0.98	1.00	0.37	0.53	0.17
X_3	-0.35	0.37	1.00	0.25	0.11
X_4	-0.50	0.53	0.25	1.00	0.08
X_5	-0.18	0.17	0.11	0.08	1.00
X_6	0.07	-0.05	-0.03	-0.52	0.01
X_7	0.13	-0.11	-0.05	-0.06	-0.02
X_8	0.00	-0.02	-0.03	-0.01	0.03
Y	-0.21	0.22	0.54	-0.04	0.15
		X_6	X_7	X_8	Y
X_1		0.07	0.13	0.00	-0.21
X_2		-0.05	-0.11	-0.02	0.22
X_3		-0.03	-0.05	-0.03	0.54
X_4		-0.52	-0.06	-0.01	-0.04
X_5		0.01	-0.02	0.03	0.15
X_6		1.00	0.17	-0.37	0.02
X_7		0.17	1.00	-0.34	-0.19
X_8		-0.37	-0.34	1.00	0.48
Y		0.50	0.20	0.60	1.00

output, transshipment delay was generated and compared with the MLR and Nonlinear regression model outputs.
The ANN sample statistics (training, testing and production) were given in Table III. The Table IV presents the ANN

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

output statistics. The error in prediction was significantly low (0.006 to 0.015). The correlation coefficient was 0.93.

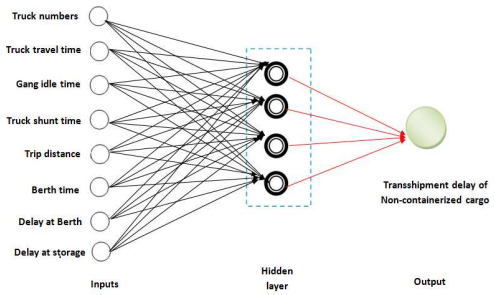


Figure 3 Hidden layer & Architecture of BPNN

TABLE III

ANN SAMPLE STATISTICS (NUMBER & PERCENTAGE)

Cargo	*Samples for Training No.*	*Samples for Testing No.*	*Samples for Prodn No.*	*Total No.*
Non-containerised	1243 (38.6 %)	638 (19.9%)	1339 (41.6%)	3221

TABLE IV

ANN MODEL PREDICTION STATISTICS

ANN output parameters	Value
R squared:	0.87
r squared:	0.87
Mean squared error:	0.001
Mean absolute error:	0.01
Correlation coefficient :	0.93

Table V
Performance of MLR & Multivariate nonlinear regression analysis

Output parameters	MLR analysis	MNLR analysis
RMS Error	8.40E-03	7.87E-02
R-Squared	0.90	0.35
Coefficient of Variation	3.90E-02	3.93E-03
Press R-Squared	0.89	0.34

B. *Multiple linear regression Models*

The multiple linear regression analysis was used to build a model between independent and dependant variables to estimate the Gross transshipment delay caused to the noncontainerized ship at Port (including delay at berth and other delays due to gang, crane and other parameters). From the multivariate correlation analysis, the correlations between the variables were found. The variables with a significant relationship have been chosen for MLR model building. The variables selected for model building were given below:

Multiple linear regression models for Gross transshipment delay of Noncontainerised cargo ships;

$$Y= 0 .108+ 3.47*10^{-05}*X_1+ 4.953*10^{-02}*X_2+0.942*X_3-1.988*10^{-02*}X_4+1.662*10^{-04}*X_5+4.397*10^{-04}*X_6+2.462*10^{-02}*X_7+1.006*X_8 \quad (1)$$

Where, X_1=Number of evacuation trucks;X_2=Truck travel time; X_3=Gang nonworking time;X_4=Truck shunting duration;X_5=Trip distance;X_6=Berth Time at berths;X_7=Waiting time at berth; X_8= other miscellaneous delays; Y=Transhipment delay.

C. *Multivariate Nonlinear regression analysis:*

Multivariate Nonlinear regression analysis was performed to build a model between independent and dependant variables to estimate the Gross transshipment delay caused to the noncontainerized category of ships. The effect of dynamics of independent variables over the dependant variables was brought in by the nonlinear analysis. The estimated MNLR model was given in eq.(2).
Nonlinear regression model:

Y = [(-9.435E-02)-(1.806E-02)*(1/SQRT(truck_Tt))+(4.51231E-03) *(1/SQRT(truck_Tt))^2+(12.41806)*(V)-(0.949)*(U)*(V)+(7.95E-02)*(V)^2 +(0.127)*(W)+(4.675489E-02)*(U)*(W)-(25.03726)*(V)*(W)+(1.599472E-02) *(W)^2+(4.856763E-02)*(X)-(0.0139986)*(U))*(X)+(1.352323)*(V)*(X) -(1.153036E-02)*(W)*(X)-(2.087984E-03)*(X)^2)] / [(1+(6.954577)*(U)*(V) +(0.3523445)*(U))*(W)-(120.3657)*(V)*(W)-(8.882952E-02)*(U))*(X) +(10.20601)*(V)*(X)+(7.149175E-03)*(W)*(X))] (2)

Where,Y = Gross transshipment delay; U = 1/√ (truck trip time);
V = (Gang idle period)^2 ;W = 1/√(truck shunting time); X = Log (craneff_ton);

V RESULTS & DISCUSSIONS

The actual service time values (observed) were plotted against artificial neural network model and MLR, MNLR forecasted outputs for Non-containerised cargo and presented in Figure 4.

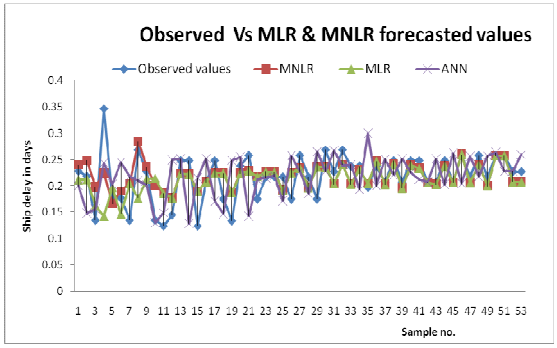


Figure 4 Observed ,MLR & MNLR ANN forecasted values

A sensitivity analysis was carried out to study the influence of port characteristics on Delays using the proposed models. The gross delay was directly proportional to the crane efficiency and truck shunting time. As the crane efficiency increase from 2000 T to 12000 T the delay might increase from 0.20 days to 0.366 days.The delay become optimised for the range of 55 to 75 shunting trucks. Also,the crane efficiency varies with the shunt trucks efficiency in transhipment. The effect got influenced by level of service or congestion levels of roads. The gross delay got affected due to port berth delays.It could be reduced by minimising the ship berth time

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

at wharf. From the sensitivity analysis, it was concluded that,even if a port well equipped port with state of the art infrastructure,may face transhipment delay, due to its operational deficiencies such as issues related to work shifts, labours discipline, insufficienct shunt trucks and cranes.

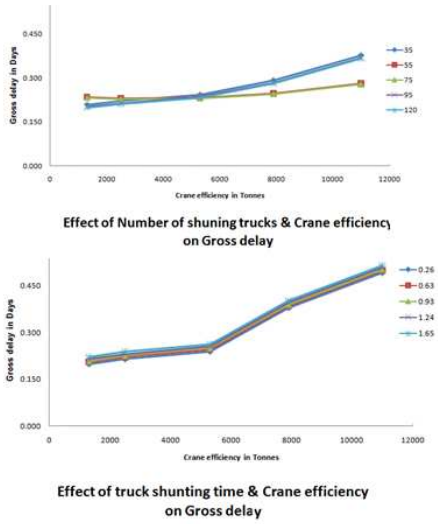


Figure 5 Sensitivity analysis outputs

VI CONCLUSION

From the outputs of ANN, MNLR and MLR analysis, it was concluded that the prediction accuracy of the ANN model was established from the R^2 (0.87) and Correlation co-efficient (0.93). This paper discussed the application of datamining techniques in predictive analysis of future delays to be faced by Non-containerised cargo at Port berths. Further, it has a scope of various issues connected with cargo transhipment in the port sector.

REFERENCES

[1] G. K. Ravi Kumar, B. Justus Rabi, Ravindra S. Hegadi, T.N. Manjunath And R. A. Archana, "Experimental study of various data masking techniques with random replacement using data volume", IJCSIS-International Journal of Computer Science and Information Security,vol. 9, No.8, August 2011, pp. 154-157.

[2] Mohammad Behrouzian Nejad, Ebrahim Behrouzian Nejad and Mehdi Sadeghzadeh, "Data Mining and its Application in Banking Industry, A Survey", IJCSIS - International Journal of Computer Science and Information Security, Vol. 9, No. 8, 2011.

[3] I. Krishna Murthy, "Data Mining- Statistics Applications: A Key to Managerial Decision Making", indiastat.com, socio - economic voices, 2010 11 pp. 1-11.

[4] A. Kusiak, "Data mining: manufacturing and service applications", International Journal of Production Research, Vol. 44, Nos. 18–19, 15th September – 1st October 2006, pp. 4175–4191.

[5] Chang Qian Gua and Rong fang (Rachel) Liu, "Modeling Gate Congestion of Marine Container Terminals-Truck Cost and Optimization," TRR, TRB No.2100, 2009, pp.58–67.

[6] Wenjuan Zhao and Anne V. Goodchild, "Impact of Truck Arrival Information on System Efficiency at Container Terminals"TRR, TRB.2162, 2010, pp. 17–24.

[7] UNCTAD Transportation newsletter, "United Nations Conference on Trade and Development,Geneva, Vol.3, 2009, pp.65-79.

[8] Nathan Huynh and Nathan Hutson, "Mining the Sources of Delay for Dray Trucks at Container Terminals", TRR, TRB 2066, 2008,pp. 41–49.

[9] U.Bugaric and D.Petrovic, "Increasing the capacity of terminal for bulk cargo unloading," Simulation modelling practice and theory, vol.15, 2007, pp. 1366–1381.

[10] Mohammed Ali, Alattan bila Varikarkae and Neelara jhsans, "Simulation of container queues for port investment decision", Proceedings of ISORA'06, Xinjiang, 2006, pp.155-167.

[11] Kasypi Mokhtar and Dr.Muhammad zaly shah, "A regression model for vessel Turnaround time", Proceedings of TAICI, 2006,pp. 10-19.

[12] Simeon Djankov, Caroline Freund and Cong S. Pham, 2006, "Trading on Time", World Bank, Development research group, 2006,pp. 1-39.

[13] F. Soriguera, D. Espinet and F. Robuste, "Optimization of the internal transport cycle in a marine container terminal managed by Straddle carriers", TRR (2006), TRB, 2007.

[14] Brian M. Lewis, Alan L. Erera and Chelsea C. White, "Impact o Temporary Seaport Closures on Freight Supply Chain Costs",TRR (2006), Vol.1963 (1), pp. 64-70

[15] Rajeev Namboothiri and Alan L. Erera, "A set partitioning heuristic for local drayage routing under time-dependent port delay",7803-8566-/04, 2004 IEEE.

[16] H. Murat Celik, "Modeling freight distribution using artificial neural networks", Transport Geography, vol.12, 2004, pp.141- 148.

[17] Peter B. Marlow and Ana C. Paixão Casaca, "Measuring lean ports Performance", Transport Management, Vol.1 (4), 2003, pp.189-202.

[18] Rahim F. Benekohal, Yoassry M. El-Zohairy and Stanley Wang, "Truck Travel Time around Weigh Stations Effects of Weigh in Motion and Automatic Vehicle Identification Systems", TRR 1716 _135,TRB 2000, pp. 138-143.

[19] Jose L. Tongzon, "Determinants of port performance and efficiency", TR,Part A, Vol.29 (3), 1995, pp.245-252.

[20] Http://www.ppiaf.org/ppiaf/sites/ppiaf.org/

[21] Ports of India website; Http://www.portsofindia.nic.in..

[22] Position paper on "The ports sector in India",Dept. of economics Affairs, Ministry of Finance, Government of India, 2009.

AUTHORS PROFILE

P.OLIVER JAYAPRAKASH is at present Assistant professor in Civil engineering department, Mepco schlenk engineering college, sivakasi, Tamilnadu. India. His field of interests includes,Soft computing applications in Freight logistics planning. He is currently pursuing his Ph.D at Anna University,Chennai under Dr.K.Gunasekaran.

Dr.K.GUNASEKARAN is an Assiociate professor in Transportation engg. Dvision of Anna University,chennai His research interests includes Simulation,Analysis and modeling of Accidents and its prevention, GIS &GPS application inTraffic analysis and management. He published several research articles in various Journals.

Dr.S.MURALIDHARAN is at present Professor in Electrical and Electronics engineering department,Mepco schlenk engineering college,sivakasi, Tamilnadu. His research interests includes Fuzzy logic and Neural Network application to power system planning,optimization and control problems. He published several research articles in varoius reputed international conferences and journals.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A DYNAMIC APPROACH FOR THE SOFTWARE QUALITY ENHANCEMENT IN SOFTWARE HOUSES THROUGH FEEDBACK SYSTEM

***Fakeeha Fatima**
Department Of Computer Science,
University OF Agriculture Faisalabad, Pakistan
Corresponding Author's
E-mail*: faatima009@gmail.com

Tasleem Mustafa
Department Of Computer Science,
University OF Agriculture Faisalabad, Pakistan

Ahsan Raza Sattar
Department Of Computer Science,
University OF Agriculture Faisalabad, Pakistan

Muhammad Inayat Khan
Department of Mathematics and Statistics,
University OF Agriculture Faisalabad, Pakistan

Waseeq-Ul-Islam Zafar
Department Of Computer Science
University OF Agriculture Faisalabad, Pakistan

ABSTRACT

Software systems are mainly changed due to changing requirements and technology which often lead to modification of software systems. In this paper dynamic approach through feedback mechanism is used to enhance the quality of the software in software houses. It involves the continual process of updating and enhancing given software by releasing new versions. These releases provide the customer with improved and error-free versions. To enhance quality VEMP (view, evaluate, maintain, performance) mechanism applied on the results gathered through feedback mechanism. By using this approach it improves overall software quality, reduce software costs, release on time and deliver software with fewer defects and get higher performance.

Keywords: Software quality, Customer Feedback, User Satisfaction, Software Quality Assurance, Dynamic Updation, Software Houses.

1.0 INTRODUCTION

The quality of a software is a major challenge in software system and is widely accepted as its conformance to customer requirements (Levin and Yadid, 2003; Vitharana and Mone, 2010). Studies indicate that 90% of all software development is maintenance and more than 50% of the total maintenance cost of software depends on rework i.e. in changing the software (Gupta *et al.*, 2010). Software systems have recently propagated greatly and become a pervasive occurrence both in the life of individuals and in culture at large. Accompanying the expansion growth of software use, it's essential to ensure the high quality of software. Sufficient software testing, authentication and error elimination are the most important techniques for improving software quality.

The main objective of this research is to produce realistic software systems that have collective and cost effective worth using an efficient development of software process to improve software quality (Martin, 2005). The quality of software could be explained by various aspects such as consistency, maintainability of the system. Dynamic approach to be use for software to enhance the quality, to improve the efficiency of programming, to reduce the cost of maintenance and promote the development of system software (Avaya *et al.*, 2007). Software developments are playing a significant role in human lives during the past years, due to the strict and vital demand of technology to make lives easier (Raz *et al.*, 2004). However, in the released software have missing functionality or errors due to the restriction of development technology, time-to-market demands and limited development resources. (Wagner, 2006; Klaus, 2010).

The cost of software problems or errors is a significant problem to global industry, not only to the producers of the software but also to their customers and end users of the software. Defects in production software can severely disrupt business operations by causing downtime, customer complaints, or errors (Wagner 2006).

1.1 RESEARCH OBJECTIVE

Software manufacturing is the methodological approach toward the expansion and preservation of the software. It had a significant impact on future of the discipline by focusing its efforts on enhancing the software quality. The primary objective of this research is the construction of programs that meet stipulation and evidently perfect, developed with in scheduled time and agreed budget. The purpose to do this research is to discover the requirements according to the changing needs of user's

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

environment that help to improve quality of system. By using dynamic approach we can upgrade a system according to the need of the user to enhance and improve software quality and make them more reliable. The online feedback mechanism is used to take responses of users.

1.2 SOFTWARE QUALITY THROUGH DYNAMIC UPDATION

Dynamic updation is a type of software development that upgrades a running system without disruption (Gorakavi, 2009; Orso et al., 2002). Software system are continually varying and developing in order to eradicate faults, enhance the performance or consistency and append better functionality to acquire better quality of the working system. Typically software updation process consists of stopping the system to be updated, performing the updation of the code and feathers and restarting the system (Taylor and Ford 2006; Chen, *et al.*, 2006). This situation is worst and take a time to maintain quality of the software (Chen and Dagnat 2011).

A essential aspect of quality is that it's not complimentary and it constantly entail efforts characteristically in reviewing, testing, examination etc. which outlay extra but on the other hand it forever append some assessment to the customer (Chen and Dagna, 2011). A general view of quality is the totality of features and characteristics of a product or service to satisfy specified or implied needs.

In this research the quality of software products enhanced during process for continuously development which involves the management control, coordination, and feedback from various contemporaneous processes during the software life cycle development and its implementation process for fault exposure, to the elimination and anticipation and the quality expansion process (Lai *et al.*, 2011; Levin and Yadid, 2003) . The excellence of software is believed to be elevated higher if it meets the standards and procedures according to the needs of the users required for the product. Software intensive companies experience re-appearing problems as well as problems due to lack of knowledge about certain technology, methods and no proper communication with the customers (Dingsoyr and Conradi, 2000). A way to reduce such problems is to make better feedback structures for a company i.e. try to learn from past successes and mistakes to improve the development process.

1.3 SOFTWARE QUALITY

Quality is a perception that requires a comprehensive and concise meaning and consequently it's difficult to measure accurately, evaluate among various services, business, and possessions (Wysocki, 2006). The high quality of the software is defined as software having no mistake and deficiencies. It's extremely hard to demonstrate that the software doesn't contain any errors. Consequently the good quality of software is not including any mistake and insufficiency. It's generally accepted that the development of high-quality software is an important challenge to the industry (Klaus, 2010). Quality is progressively more perceived as a considerable characteristic of software, Software possession, expansion, preservation and process organizations tackle with these swing are universal, not any sufficiently operation to contract through it. (Abran *et al.*, 2004; Chen and Dagnat, 2011).

1.4 ROLE OF SOFTWARE HOUSES TO GAIN SOFTWARE QUALITY

Software houses are captivating steps towards the accomplishment of quality organization system (QOS) and attaining certifications to global quality principles (Akinola, 2011). The quality of the software is a positional motivation to enhance the company's representation, attract innovation of employee and assist to remain the staff turnover low (Hellens, 2007). The software houses handled various Software Projects and the duration of each project varied from time to time depending on the scope and user requirement elicitation. Majority of the firms complained that customer don't identify what they desire until they see it and thus effecting project duration. Mostly the users know what they want but they cannot explain their requirements effectively (Olalekan *et al.*, 2011). The modifications have to be tracked, investigated and submissive to make sure elevated quality in the outcome (Khokharet *et al.*, 2010). A qualified software house usually consists of at least three enthusiastic subordinate terms (Haiwen *et al.*, 1999): business analysts who describe the business requirements of the marketplace, software expensive / programmers who generate the technological requirement and develop the software, software testers who are accountable for the entire procedure of quality administration.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

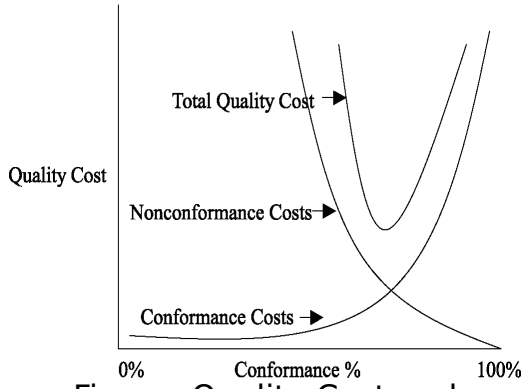


Figure: Quality Cost and Conformance Level

2.0 APPLICATION METHODOLOGY

Dynamic software updating (DSU) is a method in which a running program can be updated with innovative convention and data without interrupting its execution which must provide continuous service to fix bugs and add new features (Stoyle *et al.*, 2007). Dynamic software updating is also useful for avoiding the need to stop and start a system every time it must be patched.

2.1 Feedback mechanism in Software Houses

In this research the basic purpose is to eliminate problems and difficulties of the business customers because of the varying demand of the users need to maintain the he quality of the system. For his purpose a dynamic updation process through feedback mechanic is used to get the latest demands of the users and find bugs occur during the working (Contributor, 2006). Problems occur due to lack of knowledge about certain technology, methods and improper communication with the customers (Dingsoyr and Conradi 2000). Feedback is a significant ingredient to measure the performance of system (Akinola, 2011; Avaya *et al.*, 2007). Feedback is taken from customer through online mechanism, interviews, survey, meetings to the user who handle the system. After making changes new version is released with additional features that fulfil the current requirements of the users. A collective feedback is taken of the whole software projects.

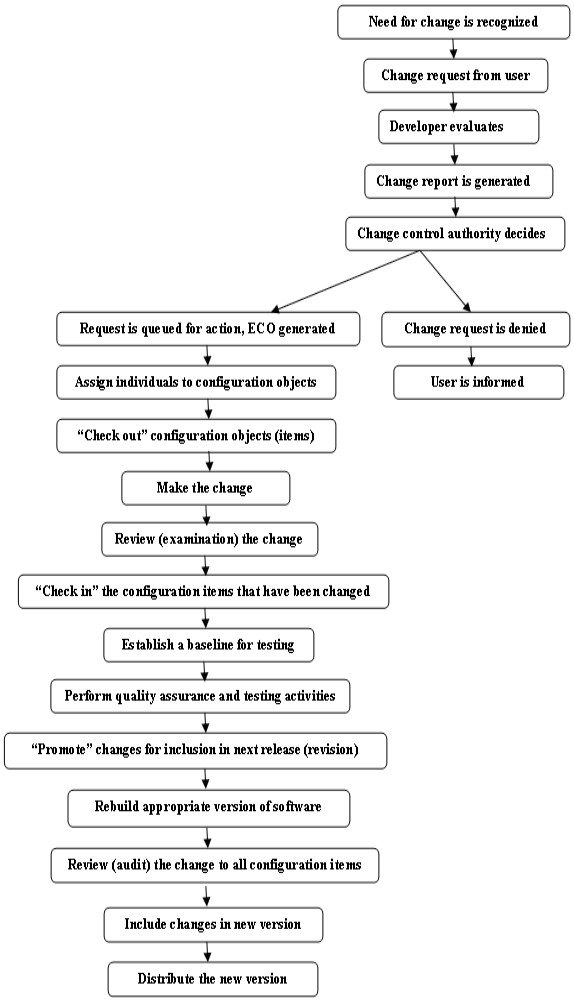


Figure : Change Control Process, (Source: Pressman, 2001)

2.2 BETA VERSION:

A beta version is launched by a corporation to releases their software or manufactured goods on a trial basis to acquire user's opinion and to investigate faults or mistake that might require to be improved. Furthermore, it gives awareness to enhance consciousness to potential customers by giving them an opportunity to "first try before you buy". A beta version is offered to the organization to check the needs and find errors in the previous version while adding new features that help to maintain the system quality and enhanced functionality.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

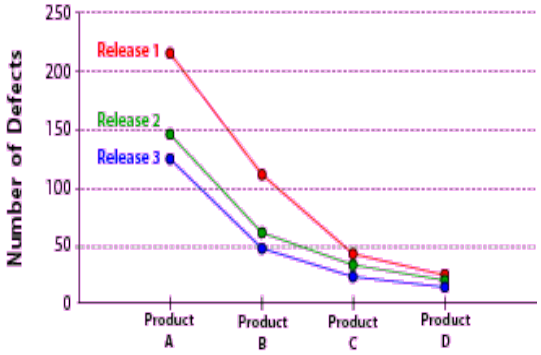


Figure: Defect rate Software Product Release

2.3 QUALITY INDICATORS

Quality benefits of software product lines can be measured in two ways. The first is how well each product matches the needs of each customer. The mass customization capabilities of software product lines directly address to measure quality (Hevner, 1997). The second is the rate of defects found in project, which can also be significantly improved by software product (Martin, 2005). The satisfied customers provide a continuing revenue stream and provide positive recommendations (Huan *et al.*, 2008).The suggested indicators are:

- Quality in feedback mechanism
- Testing process well defined
- Experienced feed backing staff

The process quality and the indicator values are judged on a five-point scale from 'very low' to 'very high' the judgement relative to the norm for the developed environment (Neil and Fenton, 2007; Akinola, 2011) . To set up the indicators, an expert judge its 'strength' as an indicator of the underlying quality attributes.

3.0 DISCUSSION

Software evolves to fix bugs and add features, but stopping and restarting existing programs to take advantage of these changes can be inconvenient and costly. Dynamic software updating (DSU) addresses these problems by updating programs while they run (Chen and Dagnat, 2011). The challenge is to develop Dynamic software updating infrastructure that is flexible, safe and efficient. Dynamic software updating enable updates that are likely to occur in practice and updated programs should be as reliable and efficient.

Feedback is an integral part of the improving a process in the software industry. Through our personalized fast quality feedback we succeeded in increasing motivation and confidence. (George, 2003). To enhance quality VEMP (view, evaluate, maintain, perform) mechanism is applied on the results gathered through feedback. By using this approach it improves overall software quality, reduce software costs, release on time and deliver software with fewer defects and get higher performance. The quality of software is the variation of software excellence at its release time and consequent efforts to manage the software throughout their functional life (Momoh and Ruhe, 2005). The protection refers to the actions that edit the software after release in the direction to get better performance and other quality features, to be adapted the product in changed situations (Wagner, 2006). Lacking of maintenance, software is in hazard of rapidly flattering obsolete. The ultimate goal of these techniques and methods are to help software developers to produce quality softwares in an economic and timely fashion.

CONCLUSION:

The consequences of the research demonstrate that dynamic technique through feedback mechanism successfully applied to improve excellence of the software by means of slight operating cost, less execution time and program volume during project development and maintenance.

Firstly, the fault reported in the preceding version eradicated. Secondly, software developers find out the requirements from user's anticipation, evaluation, complaints and then combine them what they have learnt with their strength during the research and development. Thirdly, the new features are added and remove bugs that are detected in the preceding version to get a more reliable system.

The respondent errors and suggestions help to acquired requirements from different point of view, which help better understanding of system. Enhancements in software processes would improve software quality; reduce expenditure and in time release. The common goals are to deliver the project in time and within finances.

After congregating requirements as well as information's regarding developed system a possibility revise determination would be done. The proposed work premeditated by taking the inclusive study of the accessible system. It is a system in which electronic data processing methods are used to make it error free. New techniques and procedures resolve the problems of projected system. The proposed research is relatively comprehensive and it covers all features in detailed.

REFERENCES

Akinola O., F. Ajao and O. B. Akinkunmi 2011. An Empirical Study of Software Project Management among Some Selected Software Houses in Nigeria, International Journal of Computer Science and Information Security.9(3):263-271.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Avaya I., N.Repenning, J. Sterman, M. Cusumano, and D. Ford 2007. Dynamics of Concurrent Software Development. development projects System Dynamics Review 22(1): 51-71.

Chen H.B, R.Chen, F.Z. Chen, B.Y Zhang and P.C. Yew. 2006. Live Updating Operating Systems Using Virtualization. 2nd Int'l Conf on Virtual Execution Environments (VEE). Pp: 35-44.

Chen X. and F. Dagnat 2011. Dynamic Reconfiguration on Java Internship bibliography. PP:14.

Contributor M. 2006. Integrate user feedback into new software product releases. development projects. System Dynamics Review 22(1): 51-71.

Dingsoyr T. and R Conradi 2000. Knowledge Management Systems as a Feedback Mechanism in Software Development Processes: A Search for Success Criteria. Norwegian University of Science and Technology (NTNU),Pp:15.

Fenton N., M. Neil, W. Marsh, P. Hearty and Ł. Radlinski. 2007. "Project Data Incorporating Qualitative Factors for Improved Software Defect Prediction" ICSEW '07 Proceedings of the 29th International Conference on Software Engineering Workshops IEEE Computer Society

Gorakavi P.K. 2009 "Build Your Project Using Dynamic System Development Method" published December 2009 at www.asapm.org.

Gupta C., Y Singh, and D. S. Chauhan 2010. A Dynamic Approach to Estimate Change Impact using Type of Change Propagation, Journal of Information Processing Systems, Vol.6 (4): 597.

Haiwen L, M. Ross, G.King, G. Staples and M. Jing 1999. Quality Approaches in a Large Software House, Software Quality Journal – SQJ, vol. 8(1): 21-35.

Hellens L. V. 2007. Quality Management Systems in Australian Software Houses: some problems of sustaining creativity in the software process. Australasian Journal of Information Systems,Vol 3(1):14-24.

Huan L., H. Beibei, and L. Jinhu. 2008. Dynamical Evolution Analysis of the Object-Oriented Software Systems, IEEE Congress on Evolutionary Computation, Pp:30-40.

Khokhar, N. M., A.Mansoor, S.U. Rehman, and A. Rauf 2010. MECA: Software process improvement for small organizations, Information and Emerging Technologies (ICIET), International Conference Vol. (32):1 – 6.

Klaus L. 2010. Engineering Quality Requirements Using Quality Models. ICECCS pp.245-24

Lai R., M. Garg, P. K. Kapur and S Liu 2011. A Study of When to Release a Software Product from the Perspective of Software Reliability Models, JOURNAL OF SOFTWARE, VOL. 6(4):651:661.

Levin K.D and O Yadid 2003. Optimal release time of improved versions of software packages, Faculty of Industrial Engineering and Management, Technion, Israeli Institute of Technology, Haifa 32000, Israel, Information and Software Technology, 32(1):65–70.

Martin M. 2005. Quality Improvement in Volunteer Free Software Projects: Exploring the Impact of Release Management, University of Cambridge Proceedings of the First International Conference on Open Source Systems Genova, Marco Scotto and Giancarlo Succi (Eds.): 309-310.

Orso A, Rao A, and M J Harrold 2002. A Technique for Dynamic Updating of Java Software, Proceedings of the International Conference on Software Maintenance (ICSM'02), IEEE, pp:649-658.

Pressman R S.2001. Software Engineering: a Practitioner's Approach. 5th Edition. McGraw-Hill. BOOK CO, Singapore. PP: 860.

Raz O.B, M.Shaw, P.Koopman, and C. Faloutsos. 2004. Automated Assistance for Eliciting User Expectations. School of Computer Science, Civil Engineering Department, ECE Department Carnegie Mellon University, Pittsburgh USA. 18(4): 291-300.

Stoyle G., M.Hicks, G. Bierman, P. Sewell, and I. Neamtiu, 2007. Mutatis mutandis: Safe and predictable dynamic software updating. ACM Trans. Program. Lang. Syst. Article 22 29(4):70.

Taylor, T. and D. N. Ford 2006. Tipping point failure and robustness in single development projects, System dynamics review. 22(1):51-71.

Vitharana P. and M. A. Mone 2010. Software Quality Management: Measurement and Research Directions, Syracuse University, USA. IGI Global.

Wagner. S 2006. A Literature Survey of the Software Quality Economics of Defect-Detection Techniques. In Proc. 5th ACM-IEEE International Symposium on Empirical Software Engineering. Pp: 73 – 84.

Wysocki R. K. 2006. Effective Software Project Management. John Wiley & Sons.

A SECURED CHAT SYSTEM WITH AUTHENTICATION TECHNIQUE AS RSA DIGITAL SIGNATURE

[1]Oyinloye O.Elohor [2]Ogemuno Emamuzo
/Achievers University
[1,2]Computer and Information system Achievers University
Achievers University, AUO
Owo, Ondo state, Nigeria
[1]rukkivie@yahoo.com, [2]cmcmamus@yahoo.com

[3]Akinbohun Folake [4]Ayodeji .I. Fasiku
[3]Department of Computer Science, Owo Rufus Giwa Polythenic, Owo, Ondo, Nigeria.

[4]Department of Computer Science, Federal University of Technology, Akure, Nigeria

[3]folakeakinbohun@yahoo.com,[4]Iretiayous76@yahoo.com

Abstract Over the years chat system which is an application or tool used for communicating between two or more persons over a network, has been faced with issues of security, data integrity and confidentiality of information/data, the attacks include social engineering or poisoned URL (universal resource locator). An effective attack using a poisoned URL may affect lots of users within a short period of time, since each user is regarded as a trusted user, other are plain text attack which makes communication vulnerable to eavesdropping, instant messaging client software often requires users to expose open user datagram protocol ports increasing the threat posed. The purpose of this research is to develop a secured chat system environment using Digital Signature, the digital signature is used to establish a secure communication channel, providing an improved secured technique for authentication of chat communication.

Keywords-Secure Chat System, RSA, Public modulus, public exponent, Private exponent, Private modulus, digital Signing, Verification, Communication Instant Messengers (IM)

I. INTRODUCTION

Chat system is a real-time direct text-based instant messaging communication system between two or more people using personal computers or other devices, running the same application simultaneously over the internet or other types of networks. Chat is most commonly used for social interaction, for example, people might use chat to discuss topics of shared interest or to meet other people with similar interests, businesses and educational institutions are increasingly using chat as well for example, some companies hold large online chat meetings to tell employees about new business developments, small workgroups within a company may use chat to coordinate their work [1]. In education, teachers use chat to help students practice language skills and to provide mentoring to students. More advanced instant messaging software clients also allow enhanced modes of communication, such as live voice or video calling. Online chat and instant messaging differs from other technologies such as e-mail, due to the perceived synchronicity of the communications by the users.

Instant messengers are faced with several security problems which affects the integrity, confidentiality of the data communicated, which are Denial of service attack, identity issues, privacy issues, transfer of malware through file transfer, as a worm propagator vector, poisoned URL, social engineering attack etc.

Several techniques have been employed to the transport layers (communication channel) which include TLSSSL (8). The vulnerability in the transport layer security protocol allows man-in-the-middle attackers to surreptitiously introduce text at the beginning of an SSL session, says Marsh Ray (), recent research has shown that those techniques have been diagnosed to have salient flaws, Related to Instant Messenger (IM) security, a modified Diffie-Hellman protocol suitable to instant messaging has been designed by Kikuchi et al. [2], primarily intended to secure message confidentiality against IM servers. It does not ensure authentication and also has problems similar to the IMSecure3 solutions. Most chat systems have no form of security of the communicated data. This research provides a tool for securing data in chat system. The secured chat system is designed to provide security, confidentiality, and integrity of communication between

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

parties involved by using the underlining technologies of Rivest-Shamir-Adelman (RSA) algorithm digital signature technique as its method of authentication and verification of users' .The digital signature uniquely identifies the signer of the document or message.

OPERATION OF INSTANT MESSENGERS

To conduct a conversation using instant messaging, the users must first install a compatible instant messaging program on his/her computer. On successful installation, the users are presented with a customized window from which both users will exchange other named information for effective communication. The delivery of information to the user is dependent on the availability of the user on online. Typically, IM software requires a central server which relays messages between clients. The client software allows users to maintain a list of contacts that he wants to communicate with, information transferred is via text-based communications and communication with other clients is by double clicking on the clients' detail in the contact list. The message contains the IP address of the server, the username, password and IP address of the client.When the ISP connects with the specific server, it delivers the information from the clients end of the IM software. The server takes the information and logs the user on to the messenger service, the servers locate others on the user's contact list if they are logged on to the messenger server. The connection between the PC, ISP and the messenger server stays open until the IM is closed, as illustrated in fig. 1.

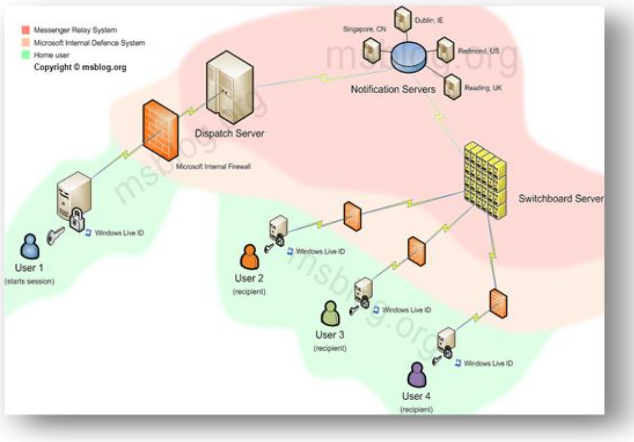


Fig 1: A windows Chat System

OVERVIEW OF EXISITNG INSTANT MESSENGERS

All Instant Messengers (IM) are categorized into five types:

Single-Protocols IMs: The five most popular IMs, based on total users, fall under the category of single-protocol IMs. In these clients connect their users often to only one or two networks of IM users, limiting contact to only those respective networks of IM users. E.g. ICQ Messenger, Skype, Yahoo IM, Windows Live Messenger, Google-Talk (Gtalk), hence single-protocol IM clients offer limited access[7].

Multi-Protocol IMs: While single-protocol IM clients offer limited access, the possibilities are endless with multi-protocol IMs. Multi-protocol IM clients allow users to connect all your IM accounts with one single chat client. The end result is a more efficient IM experience with multi-protocol IMs than using several IM clients at once. E.g; Adium, Digsby,AOL(American Online) IM, ebuddy, nimbuzz, Miranda IM, Pidgin, Yahoo IM, Windows Live Messenger. [7].

Web-Based Protocol IMs : When you cannot download an IM client web messengers are a great web-based alternative for keeping in touch with other users, unlike other multi-protocol IM clients, web messengers require nothing more than a screen name to your favorite IM and a web browser. Examples are; meebo, AIM Express Web Messenger, IM+ Web Messenger. [7].

Enterprise Protocol IMs: Instant messaging is a brilliant way to keep in touch with other users, IM is finding new-found application as a commerce-building tool in today's workplace. In addition to opening lines of communication between departments and associates throughout a company, instant messaging has helped in streamlining customer service. E.g. 24im, AIM-Pro, Big Ant, Bitwise Professional, Brosix. [7].

Portable Protocol IMs: While users cannot always download IMs to computers at work or school because of administrative control, they can utilize portable apps for IM by downloading and installing them to a USB drive; once installed, the portable apps can be run from the USB drive connecting users to all their favorite IM contacts. Examples of this protocol are; Pidgin Portable, Miranda Portable, pixaMSN, TerraIM, MiniAIM. [7].

SECURITY THREATS OF INSTANT MESSENGERS

Denial of Service (DoS)- DoS attacks can be launched in many different ways. Some may simply crash the messaging client repeatedly. Attackers may use the client to process CPU and/or memory intensive work that will lead to an unresponsive or crashed system. Flooding with unwanted messages is particularly easy when users choose to receive messages from everyone. In this case, attackers may also send spam messages such as advertisements.

Impersonation- Attackers may impersonate valid users in at least two different ways. If a user's password is captured, attackers can use automated scripts to impersonate the victim to users in his/her contact list [3]. Alternatively, attackers can seize client-to-server connections (e.g. by spoofing sequence numbers).

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IM as a Worm Propagation Vector- Here we use a broad definition of worms [4]. Worms can easily propagate through instant messaging networks using the file transfer feature. Generally, users are unsuspecting when receiving a file from a known contact. Worms successfully use this behavior by impersonating the sender. This is becoming a serious problem, as common anti-virus tools do not generally monitor IM traffic.

DNS Spoofing to Setup Rogue IM Server- Trojans like QHosts-125 can be used to modify the TCP/IP settings in a victim's system to point to a different DNS server. Malicious hackers can set up an IM server and use DNS spoofing so that victims' systems connect to the rogue server instead of a legitimate one. IM clients presently have no way to verify whether they are talking to legitimate servers. Servers verify a client's identity by checking the user name and password hash. This server-side only authentication mechanism can be targeted for IM man-in-the-middle attacks where a rogue server may pose as a legitimate server [5]. Account-related information collection, eavesdropping, impersonation and many other attacks are possible if this attack is successful.

Plaintext Registry and Message Archiving.-There are many security related settings in IM clients. Knowledgeable users can set privacy and security settings for their needs. IM clients save these settings in the Windows registry. Any technically inclined Windows user can read registry values and users with administrative power can modify those as well. Some security related IM settings saved in the registry are: encrypted password, user name, whether to scan incoming files for viruses and the anti-virus software path, whether permission is required to be added in someone's contact list, who may contact the user (only from contacts or everyone), whether to share files with others, shared directory path, and whether to ask for a password when changing security related settings. MSN Messenger even stores a user's contact list, block list and allow list in the registry[6] in a human-readable format. Attackers can use Trojan horses to modify or collect these settings with little effort. Modifying the registry may help the intruder bypass some security options like add contact authorization, file transfer permission etc. By collecting user names and password hashes, attackers can take control of user accounts. Also, the plaintext password can be extracted from the encrypted password stored in the registry using tools such as Elcomsoft's Advanced Instant Messengers Password Recovery [6]

IMPLEMENTATION OF THE SECURED CHAT SYSTEM

The secured chat system is a two-tier architecture, which offers an improvement to existing chat system which have problems of data security, denial of service attacks by providing a cheaper but secured authentication technique for chat systems. . An existing chat system model was combined with the digital signature; the system uses RSA digital signature scheme as its method of authentication. The digital signature is formed by appending to a message a set of existing private key system generated and verifiable by only that user who has formed a non-repudiated connection with the sender. The receiver and the sender are presented with several components for the establishment of a secured connection illustrated in fig 3.

MATHEMATICAL MODEL FOR THE DIGITAL SIGNATURE AUTHENTICATION OF THE SYSTEM

The users on enrolment are made to create an account which is stored in an array-linked list hash table database located at the server end of the system; the registration is completed when a user provides a username and generates the private key modulus and exponent generated from equation 1, 2, 3

$$N = p \times q \tag{1}$$

$$512 < e < \varphi(N) \tag{2}$$

Where p is the set $512 \leq p \leq 1024$ and $512 \leq q \leq p$

$$\varphi(N) = (p-1)(q-1) \tag{3}$$

The modulus and exponent is used to perform the signature operation shown in equation 4 at the request for private communication by a client

$$C = (M^e mod N) \tag{4}$$

The receiver must also establish a private connection by generating his private and public keys respectively. The message sent by the user is encrypted using the senders private key and is only decrypted using the senders public key, thus for the original message to reach the receiver, the receiver and the sender must have established a two way handshake protocol of their public keys and the verification of the process is given by the equation 5

$$M = C^d mod\ N \tag{5}$$

The keys generated are computer generated in 512 bits binary form and must be copied for signature/verification purposes.

PHASES OF THE PROPOSED SYSTEM

The phases of the system is illustrated in fig 2, it has three phases namely;

Enrolment: the system requires that the user must enroll a username, IP address and create public and private exponents and modulus which will be used for establishing a two way handshake between clients

Signature/Verification: After the enrolment phase of the system, the next phase is the signature/verification phase which involves the use of the private and the public keys/exponents. For two users to establish a secure connection, both must engage in a two way handshake procedure, they must exchange public key information when

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

they click to chat with a particular client while the client users his/her private key to certify ownership of the public key. If the verification process is not successful the user is made to reestablish the connection until successful.

Communication: This phase involves the exchange of messages between two or more users of the chat system, it requires that the users must have gone through the enrolment and the signature/verification phase before communication can be established.

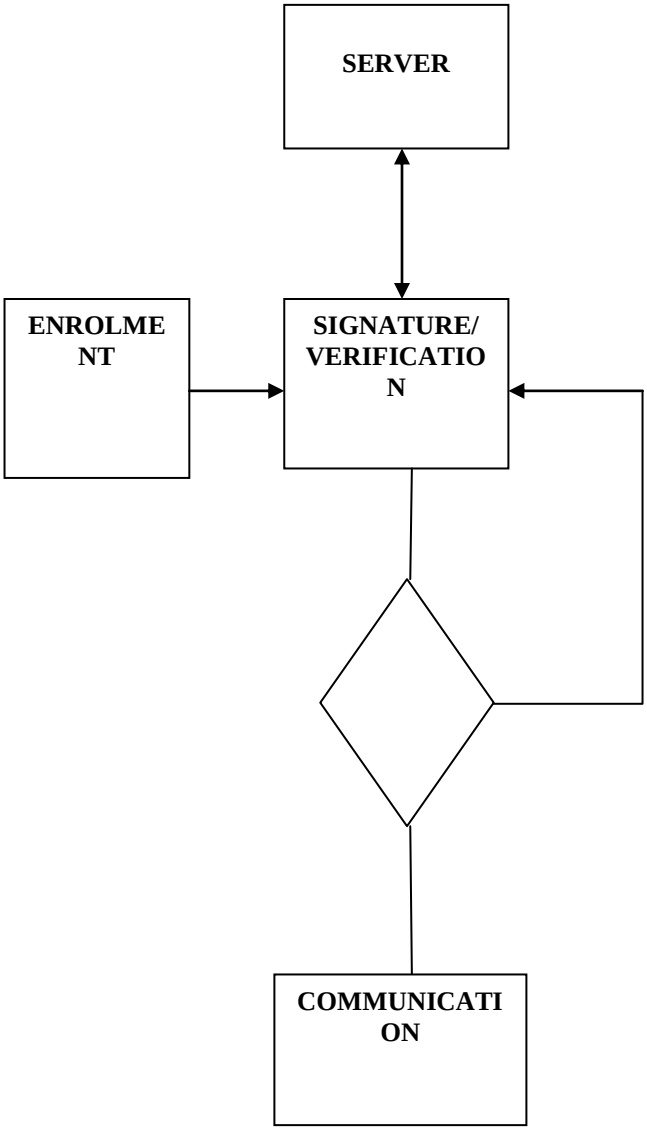


Fig 2: phases of the system

OPERATION OF THE SECURED CHAT SYSTEM

The Chat System is a Peer-to-Peer application. As shown in the fig 3, the Chat communication is achieved using XML-RPC. When a client initiates a conversation, it contacts the Chat Server to check to see the user is still actively logged in, and get the IP address and port number of the peer it wishes to communicate with. After this information is obtained, the chat session between the two peers is a client-to-client conversation and the Chat Server is no longer involved.

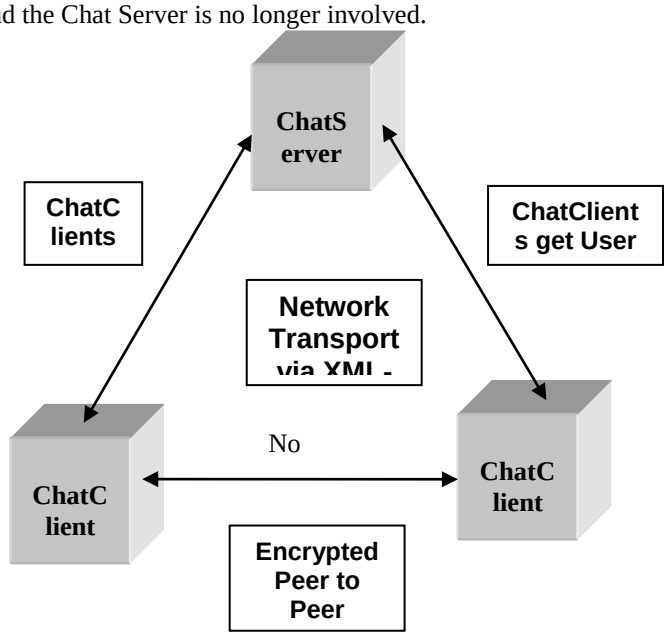


Fig 3: Operation of the secured Chat System

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

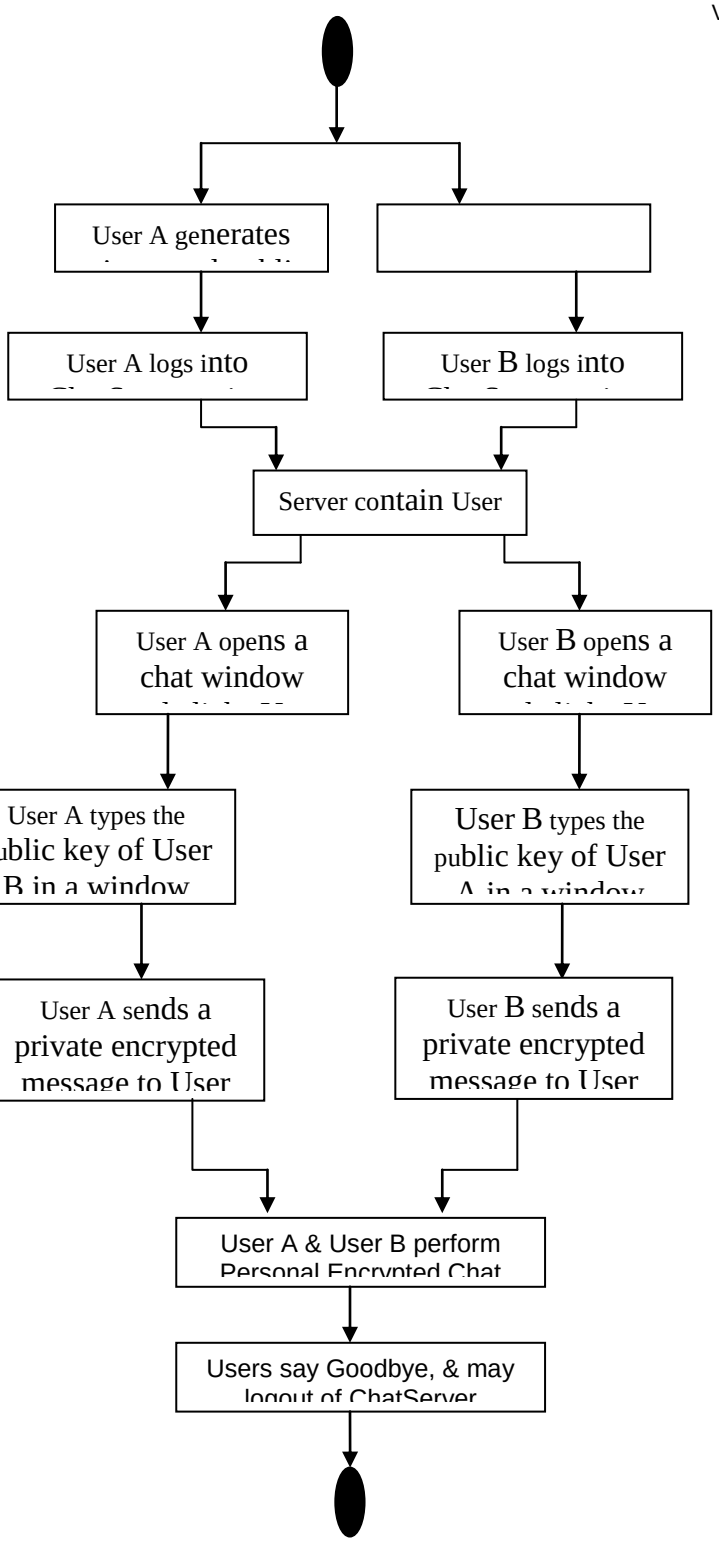


Fig 4 provides the interaction of multiple users with the Chat application, the exchange of public keys.

IMPLEMNTATION OF THE SYSTEM

The application has two broad distinctions; serverside and client side. The first step is to start the server machine, after which other users able to connect to chat system. The user is provided a window as shown in fig 5 to supply the IP address of the server system and place to enter the name to be used in the chat window.

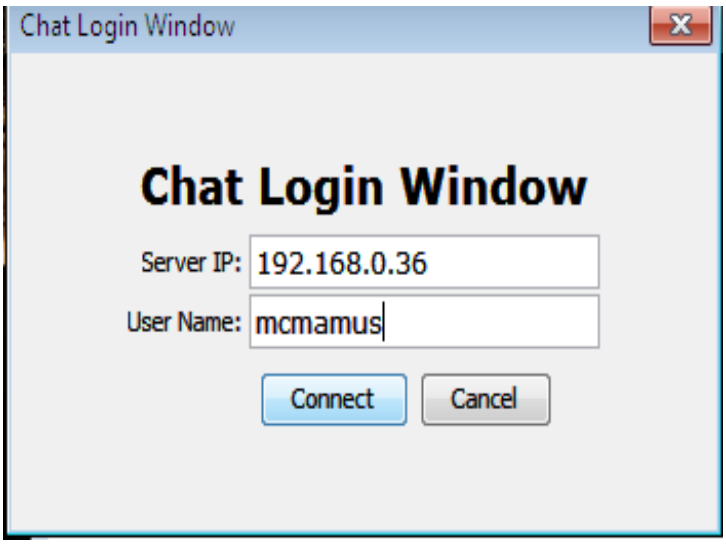


Fig 5 Login Window of The Chat System

If the server IP address is not correctly entered or the server machine is online it brings up an error message as shown in fig 6.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

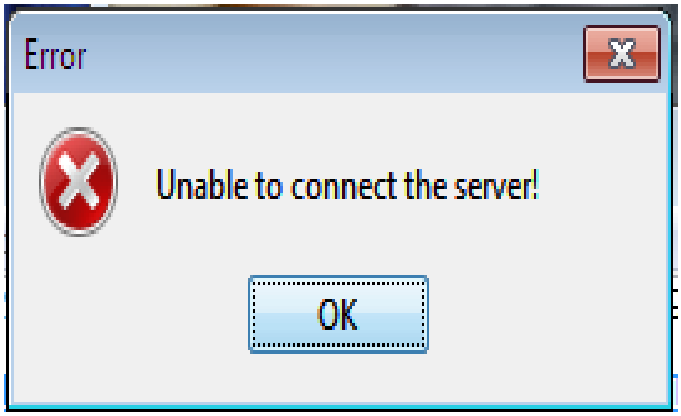


Fig 6 Error Message Dialog

The system then prompts the user to know if the user is using it for the first time or not as shown in fig 7

Fig 7 Dialog Box Showing To Know If The User Has Used The System Before Or Not

A“yes” click provides another dialog box where the user has to generate the public modulus & exponent and private modulus & exponent respectively as shown in fig 8

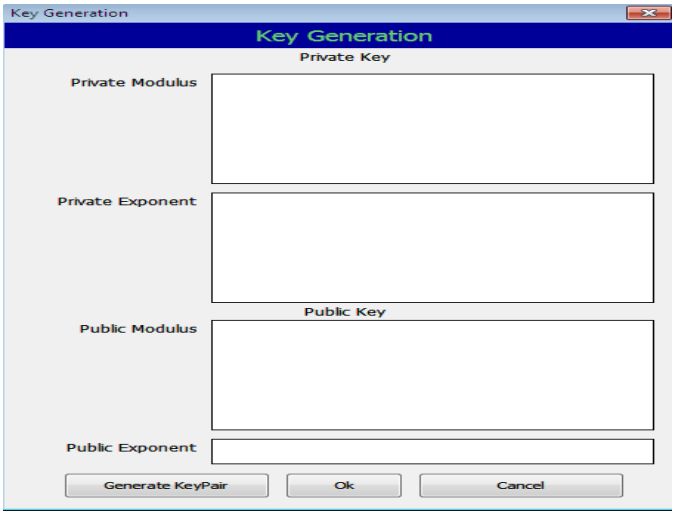


Fig 8 Key Generation

The user requires his private key to establish a private chat and he enters the public key information of the recipient, the recipient enters his private key complete the secured connection, illustrated in fig 8-12

Fig 9 Key Sign-In With Private Modulus & Exponent

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

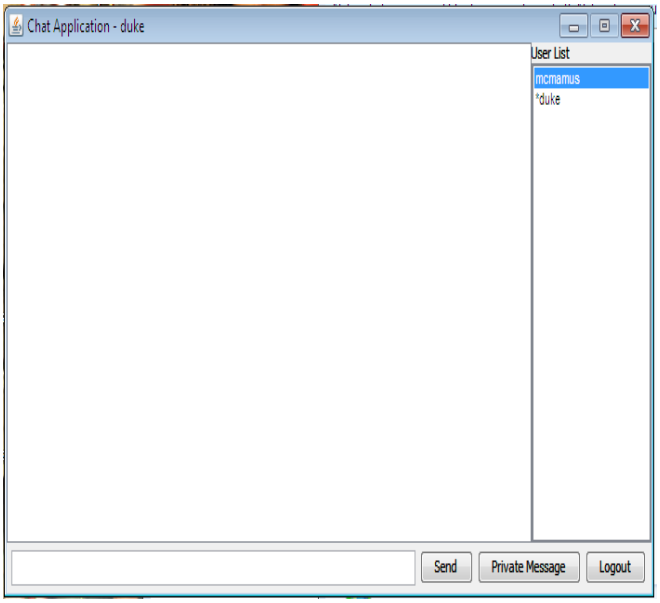


Fig 10 the Chat Window 1

When a user logs out it shows in the chat window that the user has left the chat room.

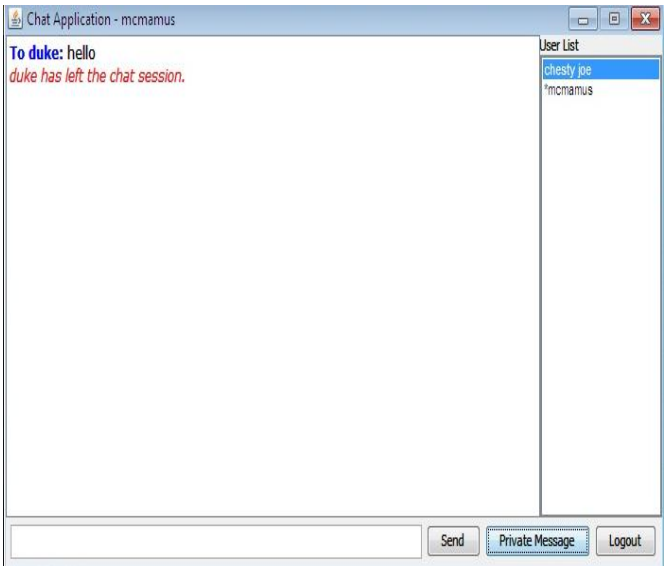


Fig 11 the Chat Window 2

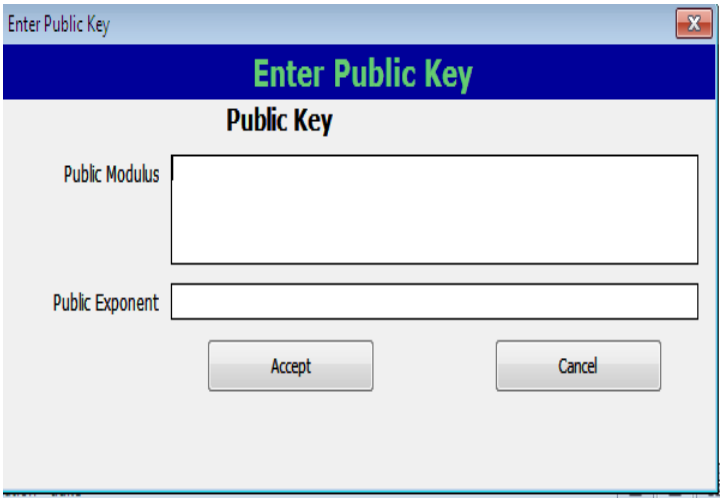


Fig 12 Public Modulus & Exponent

LIMITATIONS

The system requires the user to copy the keys and their exponent because the keys are 512 bits which makes it inconvenient and uninteresting to use.

CONCLUSION/RECOMMENDATION

Due to the efficiency and convenience of Instant Messaging (IM) communications, instant messaging systems are rapidly becoming very important tools within corporations. Unfortunately, many of the current instant messaging systems are inadequately secured and in turn are exposing users to serious security threats. In this research digital signature was used and implemented using Rivest-Shamir-Adelman (RSA) Algorithm was used in securing the chat window, and also ensuring that when a user needs to send a private message to another user of the chat system it requires that he inputs the public key of the other user, if he inputs the wrong keys the message will not be sent to the other user meaning that he is not familiar with him/her.
Further work could be done on proving a more convenient length of keys which have effective security mechanisms.

REFERENCES

[1] Bruckman, Amy S,2009,"chat(online)",Microsoft Encarta. Retrieved on 10/3/2011

[2] H. Kikuchi, M. Tada, and S. Nakanishi; 2004 "Secure instant messaging protocol preserving confidentiality against administrator," in 18th International Conference on Advanced Information Networking

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

and Applications, AINA 2004, vol. 2, Fukuoka, Japan, Mar., pp. 27–30.

[3] D. M. Kienzle and M. C. Elder,2003, "Recent worms: a survey and trends," in *Proceedings of the 2003 ACM Workshop on Rapid Malcode*. Washington, D.C., USA: ACM Press, Oct. 2003, pp. 1–10, http://pisa.ucsd.edu/worm03/worm2003-program.html [Accessed: Dec. 7, 2003].

[4] D. Petropoulos,2001, "An empirical analysis of RVP-based IM (MSN Messenger Service 3.6)," Encode Security Labs, Nov. 2001, http://www.encode-sec. com/esp0202.pdf [Accessed: Dec. 7, 2003].

[5] M. D. Murphy,2003, "Instant message security - Analysis of Cerulean Studios' Trillian application," SANS Institute, June 2003, http://www.giac.org/practical/ GSEC/Michael Murphy GSEC.pdf [Accessed: Dec. 7, 2003].

[6] D. Frase,2001, "The instant message menace: Security problems in the enterprise and some solutions," SANS Institute, Nov. 2001, http://www.sans.org/rr/ papers/60/479.pdf [Accessed: Dec. 7, 2003].

[7] Brando De Hoyos, "Instant Messaging Guide", http://www.about.com/instantmessagingguide.Retrieved on 8/4/2011

[8] Denise Doberitz (2007); Cryptographic attacks on and security flaws of SSL/TLS

AUTHORS PROFILE

Oyinloye Oghenerukevwe Elohor (Mrs.) has (MTECH.) Computer Sccience, (BSc.) In Computer Science (Technology), professional certifications in networking and a lecturer in the Department of Computer and Information Systems Achievers University, Nigeria. She is a member of IEEE. Her areas of research include Security of data, Networking and Computer Architecture.

Ogemuno E.C is a graduate of the department of Computer and Information Systems . His area of research is security programming.

Akinbohun Folake (Mrs.) has HND, PGD in computer Science, is currently running a postgraduate degree program in Computer Science. Her areas of research include computer graphics, neural networks.

Fasiku Ayodeji Ireti has a (B. Tech) in Computer Engineering, is currently running his postgraduate degree in computer Science at the Federal University of Technology, Akure, Ondo State, Nigeria. His are of research is Computer Architecture.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Motor Imagery for Mouse Automation and Control

Bedi Rajneesh Kaur
Dept. of computer engineering,
MIT COE,
Pune, India, 411038
meenubedi@hotmail.com

Bhor Rohan Tatyaba
Dept. of computer engineering,
MIT COE,
Pune, India, 411038
Rohanbhor09@yahoo.co.in

Kad Reshma Hanumant
Dept. of computer engineering,
MIT COE,Pune, India, 411038
Kad.reshma29@gmail.com

Katariya Payal Jawahar
Dept. of computer engineering,
MIT COE,
Pune, India, 411038
Payal.katariya@gmail.com

Gove Nitinkumar Rajendra
Dept. of computer engineering,
MIT COE,
Pune, India, 411038
Gove.nitinkumar@gmail.com

***Abstract*—A brain-computer interface (BCI) basically transforms the brain's electrical activity into commands that can be used to control devices such as robotic arms, pianos and other devices. With this, BCI provides a non-muscular communication channel, which can be used to help people with highly compromised motor abilities or functions. Mental imagery is the mental rehearsal of actions without overt execution. A study of motor imagery can help us to develop better neuroprosthetic systems. In this paper, we describe general concepts about motor imagery and other aspects associated with it. Recent researches in this field, has employed motor imagery in normal and brain-damaged subjects to understand the content and structure of covert processes that occur before execution of action. Finally, we propose a new system "µMAC", which will automate and control basic mouse operations using motor imagery.**

Keywords- Mu waves, Motor imagery, EEG, Neuroprosthesis, BCI, Mouse Control.

I. INTRODUCTION

Motor imagery is a one of the most studied and researched topic in the field of cognitive neuroscience. Roughly stated, motor imagery is a mental state wherein a subject imagines something. To be more specific, motor imagery is a dynamic state during which the subject mentally simulates a given action.

According to Jeannerod, motor imagery is a result of conscious access to the contents of intent of movement [1][2]. Motor imagery is a cognitive state which can be experienced virtually by anyone without more training. It is similar to many real time situations that are experienced in life like watching others performing action with intention to imitate it, making moves, imagining oneself performing action and many more [3][4]. While preparing and imagining a particular movement, the mu and central beta rhythm are desynchronized over the contralateral primary sensorimotor area [5]. This phenomenon is referred as Event-related Desynchronization (ERD)[6].

The Graz-BCI developed at Graz university of technology by the pfurtscheller's group during nineties was the firt online BCI sytem that used ERD classification in signle EEG trials to differentiate between various types of motor execution and motor imagery. After these basic studies, ERD during motor imagery has been investigated for its usability for device control by various scientists.

II. PHYSIOLOGICAL ASPECTS RELATED TO MOTOR IMAGERY

Simulating a particular activity mentally leads to activation of motor pathways. An increase is seen in muscular activity during the motor imagery [7]. During this scenario, electromyography is limited to specifically those muscles which participate in simulated action [8]. Motor imagery is independent of ability to execute the movement and is dependent on central processing mechanism.

It has been demonstrated by using various brain imaging methods that different distinct regions of cortex are activated during motor imagery i.e. MI [9]. It has been revealed in neural studies that imagined and actual actions share the same subtrates or brain areas. Various brain areas that get activated during motor imagery are supplementary motor area, primary motor cortex, the inferior parietal cortex, basal ganglia and the cerebellum.

Fig 1 shows pattern of cortical activation during mental motor imagery in normal subjects. The main Brodmann areas activated during motor imagery have been outlined on schematic views of a left hemisphere [7]. As shown in figure, there is consistent involvement of pre-motor area 6, without involvement of primary motor cortex (M1). The AC-PC line defines the horizontal reference line in magnetic resonance imaging (MRI) scan. The vertical line passing though the AC

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

(VAC) defines a verticofrontal plane. VPC is the vertical line passing through the PC [10].

The two rhythms that are strongly related with motor imagery are mu and central beta rhythms. The main characteristic that defines the mu rhythm is that it attenuates in one cerebral hemisphere during preparation of contralateral extremity movement [5], the thought of the contralateral movement or tactile electrical simulation of a contralateral limb. As these rhythms are associated with cortical areas having most direct connection with the brain's normal motor output channels, they are quite promising for BCI research.

Other thing which should be considered is that, the frequencies that are easy to be performed during ME may be too fast to imagine for a subject who is not used to motor imagery training. Due to this, most of the researchers use motor imagery with half of the velocity (0.5Hz) that are used for movement execution in simple movements [12].

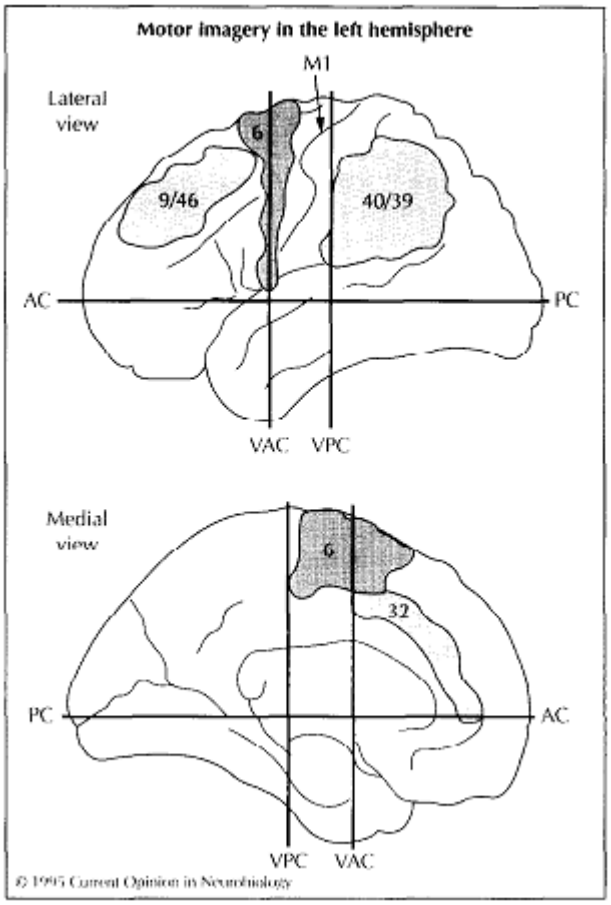


Fig.1 Pattern of cortical activation during mental motor imagery in normal subjects [7].

III. MENTAL REHEARSAL STRATEGIES FOR MOTOR IMAGERY

Basically, there are two different strategies that a subject may take or opt when asked to rehearse mentally a motor task These are -

1. Visual Imagery

2. Kinetic Imagery

1. Visual Imagery:

In this strategy, the subject produces a visual representation of their moving limb(s). The subject views himself from third person perspective (e.g. seeing one running from an external point of reference).

This type of imagery is also referred to as external imagery as for a person to view movements must have a third person perspective. VI activates regions primarily concerned with visual processing and does not obey Fitt's law nor is it correlated with excitability of the cortico-spinal path as assessed by transcranial magnetic stimulation [11].

2. Kinetic Imagery:

In this strategy, the subject rehearses or practices the particular movements using the kinesthetic feeling of the movement. Here, the subject sees himself from first person perspective. This type of imagery is also referred to as internal imagery. Each type of motor imagery has different properties with respect to both psychophysical and physiological perspectives. The motor and sensory regions that are activated during KI are same as those activated during overt movement [11].

Motor or kinesthetic imagery has to be differentiated from visual imagery because it shows different qualities: not the virtual environment is imagined in a third person's view but introspective kinesthetic feelings of moving the limb in the first person's view [10].

IV. TRAINING MOTOR SKILL

.

A subject doing mental practice/task with MI is required to have all the declarative knowledge about the various component of that specific activity/task before practicing it [13]. So, a proper training should be given to subjects about the various components of an activity/task that they are going to rehearse or practice.

The non-conscious processes involved in mental task training are best activated by the internally driven images which promote the kinesthetic feeling of movement [13]. Mental training and execution training are two complementary techniques.

According to Gandevia, motor imagery improves the dynamics of motor performance, for instance the movement trajectories [14]. The lower effect of MI training compared to ME training may be caused by lacking sensorimotor feedback which results in decreased progress in motor training in lesion patients [15]. Sufficient level of complexity of imagined motor task/activity ensures occurrence of lateralizing effect of brain activation during MI [16]. An everyday activity can

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

also be used for study of brain activations during MI in training.

This has two potential advantages [17]:

1. Easy modulation in their complexity.
2. Familiarity of task to subject helps him to generate vivid mental representation without any prior practice.

Motor imagery is widely used by athletes and musicians of improving their performance. It can be used for automation and control of mouse operations on system. Various studies have elaborated and demonstrated applications of motor imagery for controlling mouse operations [21-24].

V. THE PROPOSED SYSTEM

The systems that are proposed in these studies try to implement 1-D or 2-D control of mouse operations.

Here, we propose a system that will try to automate all the operations of mouse by using motor imagery. This includes mouse movement, left click, right click and double click. Following figure fig.2 shows a block diagram of the proposed system. Different parts of system are explained below:

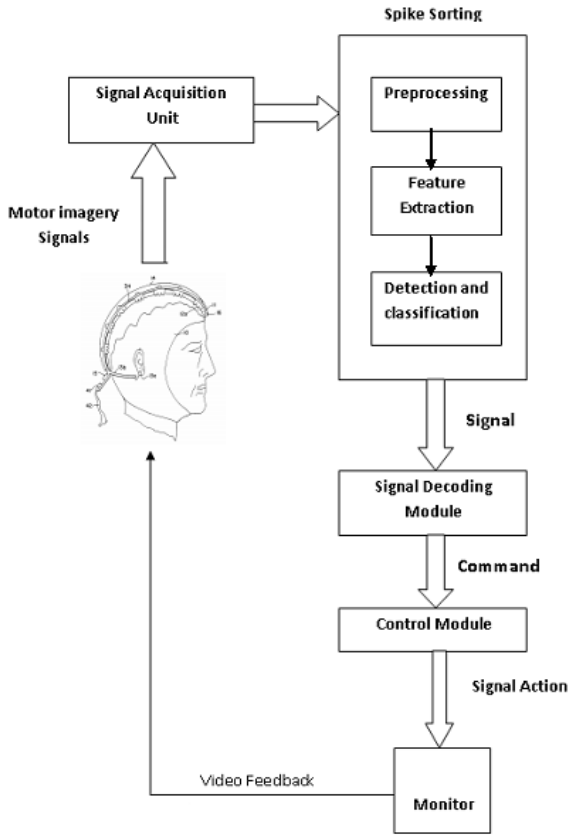


Fig 2 Block Diagram of proposed system

Signal Acquisition Unit:

The proposed system works on multi-channel EEG signals that are generated for each motor imagery activity. This unit receives the EEG signals from the sensors that are attached to the scalp of the subject's head. The signals captured by the signal acquisition unit are then passed to the spike sorting unit for further processing.

Spike Sorting Unit:

The signal captured by signal acquisition system contains noise and other unwanted spikes. These are then processed by the spike sorting unit. The signal here is processed in three phases:

a) *Preprocessing:*

This phase is responsible for artifact removal from the acquired EEG signals.

b) *Feature Extraction:*

This phase extracts differed desired features from the processed signal.

c) *Detection and classification:*

This phase is responsible for actual spike detection and its clustering into different classes.

Signal Decoding Module:

This module actually decodes/detects a particular motor imagery signal of system's concern which is further used by control module to automate the mouse operation.

Control Module:

This module on receiving the decoded signal from signal decoding module actually replicates the desired mouse operation on the monitor.

Monitor:

This is an actual display on which mouse operation is replicated.

Finally, the user receives the video feedback in the form of the mouse operation. This helps in monitoring the performance of the system.

CONCLUSION

This paper explains the basics of motor imagery, its Applications and other factors related to it. It also proposes a system for automation and control of mouse operation using brain mu and beta rhythms that are fired during this activity. This system will eventually make the existing systems more interactive and usable for physically challenged people. Apart from this, the system is quite sensitive

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

to the level of excellence with which the respective subject rehearses the desired movement or action.

In the future work, we plan to implement this system to make the proposed system usage easier and interactive for physically challenged people.

ACKNOWLEDGMENT

We are thankful to the department of computer engineering, MIT COE, for their kind support.

REFERENCES

[1] Jeannerod M. "The representing brain. Neural correlates of motor intention and imagery", Behav Sci 1994, 17:187-245
[2] Jeannerod M. "Mental imagery in the motor context", Neuropsychologia 1995, 33:1419-1432
[3] Annet J. "Motor imagery perception or action", Neuropshycologiy", 1995, 33:1395-1417,
[4] Williams JD, Rippon G., Stone BM, Annett J "Psychological correlates of dynamic imagery". Br J Psychol, 1995, 86:283-300.
[5] Pfurtscheller G., and Neuper ch., Flotzinger D. ,Pregenzer "EEG based discrimination between imagery right and left movement" Electroenceph Clin Neurophysiol 1997.
[6] Schoegl A.,Pfurtscheller G "Subject specific EEG patterns during motor imagery",19th international conference IEEE 1997
[7] Marc Jeannerod,Jean Decety "Mental motor imagery: a window into the representational statges of action" Neurobiology 1995, 5:727-732
[8] Wehner T, Bogi S , Stadler M "Task specific EMG characteristics during mental training" Psychol Res 1984 , 46:389-401
[9] A. Solodkin, P... Huichin, E chen "Small cerebral cortes." Cerebral cortex 14(2004) 1246-1255
[10] Decety J, Perani D, Jeannerod M, Bettinardi V,Tadary B, Woods R, Mazziotta JC, Fazio F "Mapping motor representation with positron emission tomography" Natur 1994, 371:600-602
[11] John Milton, Steven L. Small, Ana Solodkin "Imaging motor imagery: Methodological issues related to expertise" 2008
[12] Martin Lotze,Ulrike Halsband "Motor Imagery" 2006
[13] Jackson,P/L., Lafleur M. F. ,Maloin F.,Richards C., Doyon J., "Potential role of mental practice using motor imagery in neurologic rehabilitation" , 2001, Arch Phys Med.82:1133-114
[14] Gandevia S.C. "Mind, muscles and motor neurons" J.Sci. Med. Sport 2,167-180 1999
[15] Feltz D.L., Landers D.M. "The effect of mental practice on motor skill learning and performance – A meta analysis" J. Spot psycho 5.25-27
[16] A.J. Szameitat, S. Shen A. Starr, European journal of neuroscience 26 (2005) 3303-3308
[17] A.J. Szameitat, S. Shen , A. Starr , Neuroimage 34 (2007) 702-713
[18] Sakamoto T., Porter, Asanuma H. "Long lasting potentiation of synaptic potentials in the motor cortex produeced by simulation of the sensory cortex in the cat: basis of motor learning" 1987 Brain Res. 413, 360-364
[19] Cumming J., Hall C., "Deliberate imagery practice: the development of imagery skills in competitive athletes" 2002, J. Sports Sci. 20, 137-145
[20] Bangert M., Haeusler u. , Altenmuller E. , "On practice: how the brain connects piono keys and piono sounds" , 2001, Ann. N Y Acad. Sci 930, 425-428
[21] M. Cheng, W.Y. Jia, X.R., Gao, S.K. and F.S.Yang "mu rhythm based mouse control: an offline analysis", 2004, Clinical Neurophysiology, vol 115 pp, 745-751
[22] G.E. Fabianim D.J. McFarland, J.R. Wolpaw and G. Pfurtscheller," conversion of EEG activity into cursor movement by a brain computer interface (BCI)" IEEE trans. Neural systems and rehabilitation engineering . vol 12 (3) pp ,331-338 ,2004
[23] L.j. Trejo, r. Rosipal, B. Matthews "Brain computer interfaces for 1-D and 2-D cursor control: designs using volitical control of the EEg spectrum or steady state visual evoked potentials" IEEE trans. Neural systems and rehabilitation engineering, Vol 14 no 4, 331- 354, 2006
[24] Yuanqing Li , Jinyi long, Tianyou Yu, Zhuliang Yu, Chuanchu Wang, Haihong Zhang, Cuntai Guan "An EEG based BCI system for 2-D cursor control by combining Mu/Beta rhythm and P300 potential" , 2010, IEEE trans., On Biomedical Engineering

AUTHORS PROFILE

1. Prof. Rajneesh kaur Bedi
Head of Department
Department of computer engineering
MIT COE, Pune, India
411038

2. Mr. Bhor Rohan Tatyaba
Currently pursuing bachelor's degree in computer engineering at, Department of computer engineering MIT COE, Pune, India, 411038

3. Miss. Kad Reshma Hanumant
Currently pursuing bachelors degree in computer engineering at, Department of computer engineering MIT COE, Pune, India, 411038

4. Miss. Katariya Payal Jawahar
Currently pursuing bachelors degree in computer engineering at, Department of computer engineering MIT COE, Pune, India, 411038

5. Mr. Gove Nitinkumar Rajendra
Currently pursuing bachelors degree in computer engineering at, Department of computer engineering MIT COE, Pune, India, 411038

POLICY VERFICATION, VALIDATION AND TROUBLESHOOTING IN DISTRIBUTED FIREWALLS

P.SENTHILKUMAR
Computer Science & Engineering
Affiliated to Anna University of Technology
Coimbatore, Tamilnadu, India
psenthilnandha@gmail.com

Dr.S.ARUMUGAM
CEO
Nandha Engineering College
Erode, Tamilnadu, India
dotearumugam@yahoo.co.in

***Abstract*— The Internet is one of the largest engineered systems ever deployed, has become a crucial technology for our society. It has changed the way people perform many of their daily activities from both a personal perspective and a business perspective. Unfortunately, there are risks involved when one uses the Internet. These risks coupled with advanced and evolving attack techniques place heavy burdens on security researchers and practitioners while trying to secure their networking infrastructures .Distributed firewalls are often deployed by large enterprises to filter the network traffic. Problem statement: In conventional firewall system is only verified user specified policy. But also find the inconsistencies of the firewalls. Approach: In our approach is to implement the Policy Verification, Policy Validation and Troubleshooting in Distributed Firewalls. Input: Our approach input as user specified firewall policy or security rule of the system, Administrator policy. Output: Our approach output as satisfies policy the property and troubleshooting the some problems in firewalls. In some cases the firewall cannot be work properly at the time system administrator or firewalls administrator to troubleshooting the problem.**

Keywords- Policy Verification, Policy Validation, and Troubleshooting

I. INTRODUCTION TO FIREWALL

A firewall is a program that keeps your computer safe from hackers and malicious software. The firewall is also computer hardware or software that limits access to a computer over a network or from an outside source. The firewall is used to create security check points at the boundaries of private network. [11]

The firewalls are placed at the entry points of the private network or public network. In the case of companies, if when ordinary firewall is used everyone were given the same class policy, but distributed firewalls everyone using separate policy.

The firewall is a machine or collection of machines between two networks, to meet the following criteria:

- All traffic between the two networks must pass through the firewall.
- The firewall has a mechanism to allow some traffic to pass while blocking other traffic.
- The rules describing what traffic is allowed enforce the firewall's policy.
- Resistance to security compromise.
- Auditing and accounting capabilities.
- Resource monitoring.
- No user accounts or direct user access.
- Strong authentication for proxies (e.g., smart cards rather than simple passwords).[1]

In this paper to present Policy Verification, Policy Validation, and Troubleshooting. The figure 1.1 represents the simple firewall diagram.

II. THE DISTRIBUTED FIREWALL

A distributed firewall uses a different policy, but pushes enforcement towards the edges. [2, 12, 13]

Policy

A "security policy" defines the security rules of a system. Without a defined security policy, there is no way to know what access is allowed or disallowed. The distribution of the policy can be different and varies with the implementation. It can be either directly pushed to end systems, or pulled when necessary. [2]

Policy Language

Policy is enforced by each individual host that participates in a distributed firewall. This policy file is consulted before processing incoming or outgoing messages, to verify their compliance.

III. POLICY VERIFICATION

Policy verification is enforced by the each incoming packet as per the user specified policy and also verifies the inconsistencies. The given a firewall and a set of property rules, the verification is successful

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

if and only if every property rule is satisfied by the firewall. [5]**.**

IV. POLICY VALIDATION

Firewall configurations should be validated it means checking that the configuration would enable the firewall to perform the security functions that we expect it to do and that it complies with the security policy of the organization. You cannot validate a firewall by looking at the policy alone. The policy is an indicator, but not the true state. The only way to ensure that a firewall is behaving correctly. [12] A manual validation is most effective when done as a team exercise by the security manager, firewall administrator, network architect, and everyone else who has a direct involvement in the administration and management of the organization's network security.
The policy validation system is concerned there are two distinct kinds of failure as follows [12]

Host Failure: Any of the network hosts can fail at any time. The host failure may be difficult to distinguish from a network failure, from the perspective of the rest of the network. Recovery, however, is somewhat different.

Network Failure The network can fail at any time, or can simply not be laid out as expected. These can be ignored or reported to the root Manager in some way, as they indicate a network status that the administrator ought to be made aware of. [12]

V. TROUBLESHOOTING

The troubleshooting a firewall is much an iterative problem. The failures in network programs are not limited to firewall issues. These failures may be caused by security changes. Therefore, you have to determine whether the failure is accompanied by a Windows Firewall Security Alert that indicates that a program is being blocked. [1]

Failures that are related to the default firewall configuration appear in two ways:

I. Client programs may not receive data from a server.

II. Server programs that are running on a Windows XP-based computer may not respond to client requests. For example, the following server programs may not respond.

- A Web server program, such as Internet Information Services (IIS)
- Remote Desktop
- File sharing

Troubleshooting the firewall

Follow these steps to diagnose problems:

1. To verify that TCP/IP is functioning correctly, use

the **ping** command to test the loopback address (127.0.0.1) and the assigned IP address.

2. Verify the configuration in the user interface to determine whether the firewall has been unintentionally set to **Off** or **On with No Exceptions**.

3. Use the **netsh** commands for Status and Configuration information to look for unintended settings that could be interfering with expected behavior.

4. Determine the status of the **Windows Firewall/Internet Connection Sharing** service by typing the following at a command prompt:
sc query sharedaccess

Troubleshoot service startup based on the Win32 exit code if this service does not start.

5. Determine the status of the Ipnat.sys firewall driver by typing the following at a command prompt:

sc query ipnat

This command also returns the Win32 exit code from the last start try. If the driver is not starting, use troubleshooting steps that would apply to any other driver.

6. If the driver and service are both running, and no related errors exist in the event logs, use the **Restore Defaults** option on the **Advanced** tab of **Windows Firewall** properties to eliminate any potential problem configuration.

7. If the issue is still not resolved, look for policy settings that might produce the unexpected behavior. To do this, type **GPResult /v > gpresult.txt** at the command correctly, use the **ping** command to test theprompt, and then examine the resulting text file for configured policies that are related to the firewall.

I. FIGURES AND TABLES

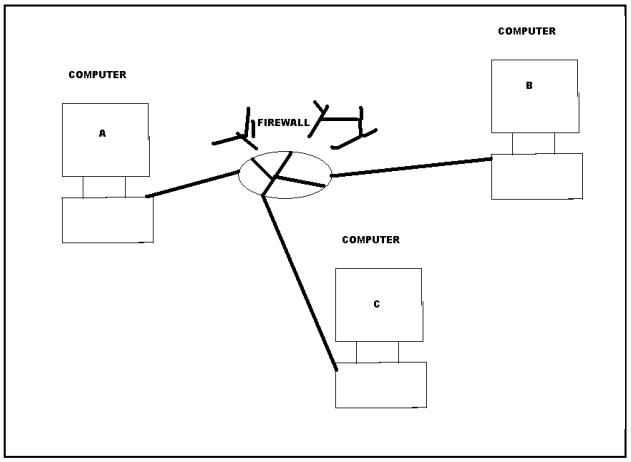


Figure1.1 Firewall Diagram

VII. RELATED WORK

Current research Policy Verification, Policy Validation and Troubleshooting in distributed firewall mainly focus the following.

1. Verifying and validating the security policy in the networks.[12]

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

2. The testing and validating firewalls regularly.[3]
3. Identify the vulnerability analysis. [11]
4.Very strong authorization and authentication for each firewalls

VIII. CONCLUSION AND FUTURE WORK

Firewalls provide proper security services if they are correctly configured and efficiently managed. Firewall policies used in enterprise networks are getting more complex as the number of firewall rules and devices becomes larger. In this paper to presented policy verification, policy validation and finding troublesome problem in the firewall.

It is an iterative process of designing a firewall. Our approach can be help to eliminate the errors in firewall policies.

ACKNOWLEDGMENT

I would like to thank my outstanding research supervisor & advisor, Dr.S.Arumugam, for his advice, support and encouragement throughout the research work. I am also very grateful to him for keeping faith in my research work even through times of slow progress.

I would like to thank my parents, brothers, sisters, and my dear V.Asha for giving me everything that I have. Your faith and love have provided me the only truth I have known in my life. Thank you.

Finally I express thanks to GOD.

REFERENCES

[1] A.X. Liu, Firewall policy verification and troubleshooting, in: Proceedings IEEE International Conference on Communications (ICC), May 2008.

[2] Al-Shaer, E. and Hazem Hamed, ''Discovery of Policy Anomalies in Distributed Firewalls,'' *Proceedings of IEEE INFOCOM'04*, March, 2004.

[3] El-Atawy, A., K. Ibrahim, H. Hamed, and E. Al- Shaer, ''Policy Segmentation for Intelligent Firewall Testing,'' *1st Workshop on Secure Network Protocols (NPSec 2005)*, November, 2005.

[4] Eppstein, D. and S .Muthukrishnan, ''Internet Packet Filter Management and Rectangle Geometry.''*Proceedings of 12th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, January,2001.

[5] Hamed, Hazem, Ehab Al-Shaer and Will Marrero,''Modeling and Verification of IPSec and VPN Security Policies,'' *Proceedings of IEEE ICNP'2005*, November, 2005.

[6] Hari, B., S. Suri, and G. Parulkar, ''Detecting and Resolving Packet Filter Conflicts.'' *Proceedings of IEEE INFOCOM'00*, March, 2000.

[7] Lakkaraju, K., R. Bearavolu, and W. Yurcik, '' N VisionIP – A Traffic Visualization Tool for Large and Complex Network Systems,'' *International Multiconference on Measurement, Modelling, and Evaluation of Computer-Communication Systems (Performance TOOLS)*, 2003.

[8] Lee, Chris P., Jason Trost, Nicholas Gibbs, Raheem Beyah, John A. Copeland, ''Visual Firewall: Real-time Network Security Monitor,'' *Proceedingsof the IEEE Workshops on Visualization for Computer Security*,p. 16, October 26-26, 2005.

[9] Liu, A. X., M. G. Gouda, H. H. Ma, and A. H. Ngu, ''Firewall Queries,'' *Proceedings of the 8th International Conference on Principles of* cited. *Distributed Systems*, LNCS 3544, T. Higashino Ed.,Springer-Verlag, December, 2004.

[10] *Tufin SecureTrack: Firewall Operations Management Solution*, http://www.tufin.com .

[11] E. Al-Shaer, H. Hamed, R. Boutaba, and M. Hasan. Conflict classification and analysis of distributed firewall policies. IEEE JSAC, 23(10), October 2005.

[12]. Kyle Wheeler. Distributed firewall policy validation, December 7, 2004.

[13] S. M. Bellovin, "Distributed Firewall",; login*:* magazine, Special issue on Security, November 1999.

AUTHORS PROFILE

1. P.Senthilkumar, is the Assistant Professor in the Department of Computer Science & Engineering, Affiliated to Anna University of Technology, Coimbatore, Tamilnadu, India. He obtained his Bachelor and Master degree in Computer Science and Engineering from Anna University, Chennai in the year 2005 and 2008 respectively. He has pursuing the Ph.D Programme at Anna University of Technology, Coimbatore. He has 6 years of Teaching Experience and authored 4 research papers in International Journals and Conferences. His current area of research includes Computer Networks, Network Security, and Firewalls Concept. He is a member of various professional societies like ISTE, International Association of Engineers, and Computer Science Teachers Association, International association of Computer Science and Information Technology and Fellow in Institution of Engineers (India). He is a reviewer and editor for various international conferences Email:psenthilnandha@gmail.com.

2**. Dr. S. ARUMUGAM**, received the PhD. Degree in Computer Science and Engineering from Anna University, Chennai in 1990.He also obtained his B.E(Electrical and Electronics Engg.) and M.Sc. (Engg) (Applied Electronics)Degrees from P.S.G College of Technology, Coimbatore, University of Madras in 1971 and 1973 respectively. He worked in the Directorate of Technical Education, Government of Tamil Nadu from 1974 at various positions from Associate Lecturer, Lecturer, Assistant Professor, Professor, Principal, and Additional Director of Technical Education. He has guided 4 PhD scholars and guiding 10 PhD scholars. He has published 70 technical papers in International and National journals and conferences. His area of interest includes network security, Biometrics and neural networks. Presently he is working as Chief Executive Officer, Nandha Engineering College Erode.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Detection and Tracking of objects in Analysing of Hyper spectral High-Resolution Imagery and Hyper spectral Video Compression

T. Arumuga Maria Devi [1]
Assistant Professor, Dept. of CITE

Nallaperumal Krishnan[2]
HOD, SMIEEE, Dept. of CITE

K.K Sherin [3]
PG Scholar

Mariadas Ronnie C.P [4]
PG Scholar

Centre for Information Technology and Engineering,
Manonmaniam Sundaranar University, Tirunelveli
[1]Email: deviececit@gmail.com.Phone No:9677996337
[2] Email: krishnan@msuniv.ac.in Phone No: 9443117397
[3]Email: sherinkk83@yahoo.com.Phone No:9442055500
[4]Email: mariadasronnie@yahoo.co.in.Phone No:8089713017

***Abstract*— This paper deals mainly with the performance study and analysis of image retrieval techniques for retrieving unrecognized objects from an image using Hyper spectral camera and high-resolution image and retrieving unrecognized objects from an image using Hyper spectral camera at low light resolution. The main work identified is that efficient retrieval of unrecognized objects in an image will be made possible using spectral analysis and spatial analysis. The methods used above to retrieve unrecognized object from a high-resolution image are found to be more efficient in comparison with the other image retrieval techniques. The detection technique to identify objects in an image is accomplished in two steps: anomaly detection based on the spectral data and the classification phase, which relies on spatial analysis. At the classification step, the detection points are projected on the high-resolution images via registration algorithms. Then each detected point is classified using linear discrimination functions and decision surfaces on spatial features. The two detection steps possess orthogonal information: spectral and spatial. The identification of moving object in a camera is not possible in a low light environment as the object has low reflectance due to lack of lights. Using Hyper spectral data cubes, each object can be identified on the basis of object luminosity. Moving object can be identified by identifying the variation in frame value. The main work identified are that efficient retrieval of unrecognized objects in an image will be made possible using Hyper spectral analysis and various other methods such as Estimation of Reflectance, Feature and mean shift tracker, Traced feature located on image, Band pass filter (Background removal) etc. These methods used above to retrieve unrecognized object from a low light resolution are found to be more efficient in comparison with the other image retrieval techniques. The objects in an image may require that its edges should be smoother in order to make it detect easily by the receiver when it is send from one machine to another. As the image and video may be needed to be send from source to destination, due to huge amount of data that may be required for processing, retrieval and storage, because of the high resolution property of images, compression is a necessity. In order to overcome the problems associated with it, Transcoding technique is used by using filter arrays and lossless compression techniques.**

***Keywords*— Anomaly suspect, spectral and spatial analysis, linear discrimination functions, registration algorithms, filter arrays mean shift algorithms, spectral detection.**

.

I. INTRODUCTION

The process of recovering unrecognized objects in an image is a trivial task which finds its need in recognizing objects from a distant location. Since there is a need in retrieving unrecognized objects from a high-resolution image, some form of object extraction method from an image is necessary. Remote sensing, for example is often used for detection of predefined targets, such as vehicles, man-made objects, or other specified objects. Since the identification of moving object in a camera is not possible from distant location, to overcome this problem we can use Hyper spectral camera to identify the object. A new technique is thus applied that combines both spectral and spatial analysis for detection and classification of such targets[4][5]. Fusion of data from two sources, a hyper spectral cube and a high-resolution image, is used as the basis of this technique. Hyper spectral images supply information about the physical properties of an object while suffering from low spatial resolution. There is another problem in a Hyper spectral image, that, it does not identify what an object is, rather, it will detect the presence of an object. In the case of a high resolution image, since the image is such that it does not show the presence of an object, some sort of mechanism is thus needed. That is why, the fusion of the two, the Hyper spectral image and the high-resolution image are used to successfully retrieve the unrecognized object from an image. The use of high-resolution images enables high-fidelity spatial analysis in addition to the spectral analysis. The detection technique to identify objects in an image is accomplished in two steps: anomaly detection based on the spectral data and the classification phase, which relies on spatial analysis. At the classification step, the detection points are projected on the high-resolution images via registration algorithms. Then each detected point is classified using linear discrimination functions and decision surfaces on spatial features. The two detection steps possess orthogonal information: spectral and spatial. At the spectral detection step, we want very high probability of detection, while at the spatial step, we reduce the number of false alarms. The problem thus relies in the area of identifying a specific area in a high-resolution image to know the presence of objects in that area. Each region selected upon the user's interest should be able to detect any presence of objects in that area. The process of recovering unrecognized objects from an image in low light is a trivial task which finds its need in recognizing objects from a distant location. Since there is a need in retrieving unrecognized objects from the image, some form of object extraction method from an image is necessary. The application of detecting objects from an image is as follows. Here, we focus on the problem of tracking objects through challenging conditions, such as tracking objects at low light where the presence of the object is difficult to identify. For example, an object which is fastly moving on a plane surface in an abrupt weather condition is normally difficult to identify. A new

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

framework that incorporates emission theory to estimate object reflectance and the mean shift algorithm to simultaneously track the object based on its reflectance spectra is proposed. The combination of spectral detection and motion prediction enables the tracker to be robust against abrupt motions, and facilitate fast convergence of the mean shift tracker. Video images are moving pictures which are sampled at frequent intervals usually, 25 frames per second and stored as sequence of frames. A problem, however, is that digital video data rates are very large, typically in the range of 150 Megabits/second. Data rates of this magnitude would consume a lot of the bandwidth in transmission, storage and computing resources in the typical personal computer. Hence, to overcome these issues, Video Compression standards have been developed and intensive research is going on to derive effective techniques to eliminate picture redundancy, allowing video information to be transmitted and stored in a compact and efficient manner. A video image consists of a time-ordered sequence of frames of still images as in figure 1. Generally, two types of image frames are defined: Intra-frames (I-frames) and Inter-frames (P- frames). I-frames are treated as independent key images and P-frames are treated as Predicted frames. An obvious solution to video compression would be predictive coding of P-frames based on previous frames and compression is made by coding the residual error. Temporal redundancy removal is included in P-frame coding, whereas I-frame coding performs only spatial redundancy removal. Related to the implementation of Transcoding, the work is as follows. The objective of this work is to study the relationship between the operational domains for prediction, according to temporal redundancies between the sequences to be encoded. Based on the motion characteristics of the inter frames, the system will adaptively select the spatial or wavelet domain for prediction. Also the work is to develop a temporal predictor which exploits the motion information among adjacent frames using extremely low side information. The proposed temporal predictor has to work without the requirement of the transmission of complete motion vector set and hence much overhead would be reduced due to the omission of motion vectors.

Spatial and Wavelet Domain: Comparison

Image compression has become increasingly of interest in both data storage and data transmission from remote acquisition platforms (satellites or airborne) because, after compression, storage space and transmission time are reduced. So, there is a need to compress the data to be transmitted in order to reduce the transmission time and effectively retrieve the data after it has been received by the receiver. In video compression, each frame is an array of pixels that must be reduced by removing redundant information. Video compression is usually done with special integrated circuits, rather than with software, to gain performance. Standard video is normally about 30 frames/sec, but 16 frames/sec is acceptable to many viewers, so frame dropping provides another form of compression. When information is removed out of a single frame, it is called *intraframe* or *spatial* compression. But video contains a lot of redundant interframe [14] information such as the background around a talking head in a news clip. Interframe compression works by first establishing a key frame that represents all the frames with similar information, and then recording only the changes that occur in each frame. The key frame is called the "I" frame and the subsequent frames that contain only "difference" information are referred to as "P" (predictive) frames. A "B" (bidirectional) frame is used when new information begins to appear in frames and contains information from previous frames and forward frames. One thing to keep in mind is that interframe compression provides high levels of compression but is difficult to edit because frame information is dispersed. Intraframe compression contains more information per frame and is easier to edit. Freeze frames during playback also have higher resolution. The aim is now to determine the operational mode of video sequence compression according to its motion characteristics. The candidate operational modes are spatial domain and wavelet domain. The wavelet domain is extensively used for compression due to its excellent energy compaction. However, it is pointed out that motion estimation in the wavelet domain might be inefficient due to shift invariant properties of wavelet transform. Hence, it is unwise to predict all kinds of video sequences in the spatial domain alone or in the wavelet domain alone. Hence a method is introduced to determine the prediction mode of a video sequence adaptively according to its temporal redundancies. The amount of temporal redundancy is estimated by the inter frame correlation coefficients of the test video sequence. The inter frame correlation coefficient between frames can be calculated. If the inter frame correlation coefficients are smaller than a predefined threshold, then the sequence is likely to be a high motion video sequence. In this case, motion compensation and coding the temporal prediction residuals in wavelet domain would be inefficient; therefore, it is wise to operate on the sequence in the spatial mode. Those sequences that have larger inter frame correlation coefficients are predicted in direct spatial domain. The frames that have more similarities with very few motion changes are coded using temporal prediction in integer wavelet domain.

Discrete Wavelet Transform (DWT)

Hyperspectral images usually have a similar global structure across components. However, different pixel intensities could exist among nearby spectral components or in the same component due to different absorption properties of the atmosphere or the material surface being imaged. This means that two kinds of correlations may be found in hyperspectral images: intraband correlation among nearby pixels in the same component, and interband correlation among pixels across adjacent components. Interband correlation should be taken into account because it allows a more compact representation of the image by packing the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

energy into fewer number of bands, enabling a higher compression performance. There are many technologies which could be applied to remove correlation across the spectral dimension, but two of them are the main approaches for hyperspectral images: the KLT and the DWT Discrete Wavelet Transform. (DWT) is the most popular transform for image-based application. They have lower computational complexity, and they provide interesting features such as component and resolution scalability and progressive transmission. A 2-dimensional wavelet transform is applied to the original image in order to decompose it into a series of filtered sub band images. At the top left of the image is a low-pass filtered version of the original and moving to the bottom right, each component contains progressively higher-frequency information that adds the detail of the image. It is clear that the higher-frequency components are relatively sparse, i.e., many of the coefficients in these components are zero or insignificant. When using a wavelet transform to describe an image, an average of the coefficients-in this case, pixels-is taken. Then the detail coefficients are calculated. Another average is taken, and more detail coefficients are calculated. This process continues until the image is completely described or the level of detail necessary to represent the image is achieved. As more detail coefficients are described, the image becomes clearer and less blocky. Once the wavelet transform is complete, a picture can be displayed at any resolution by recursively adding and subtracting the detail coefficients from a lower-resolution version. The wavelet transform is thus an efficient way of decorrelating or concentrating the important information into a few significant coefficients. The wavelet transform is particularly effective for still image compression and has been adopted as part of the JPEG 2000 standard and for still image texture coding in the MPEG-4 standard[28][30][31].

Motion Estimation Prediction

By Motion estimation, we mean the estimation of the displacement of image structures from one frame to another. Motion estimation from a sequence of images arises in many application areas, principally in scene analysis and image coding. Motion estimation obtains the motion information by finding the motion field between the reference frame and the current frame. It exploits temporal redundancy of video sequence, and, as a result, the required storage or transmission bandwidth is reduced by a factor of four. Block matching is one of the most popular and time consuming methods of motion estimation. This method compares blocks of each frame with the blocks of its next frame to compute a motion vector for each block; therefore, the next frame can be generated using the current frame and the motion vectors for each block of the frame. Block matching algorithm is one of the simplest motion estimation techniques that compare one block of the current frame with all of the blocks of the next frame to decide where the matching block is located. Considering the number of computations that has to be done for each motion vector, each frame of the video is partitioned into search windows of size H*W pixels. Each search window is then divided into smaller macro blocks of size, say, 8*8 or 16*16 pixels. To calculate the motion vectors, each block of the current frame must be compared to all of the blocks of the next frame with in the search range and the Mean Absolute Difference for each matching block is calculated. The block with the minimum value of the Mean Absolute Difference is the preferred matching block. The location of that block is the motion displacement vector for that block in current frame. The motion activities of the neighboring pixels for a specific frame are different but highly correlated since they usually characterize very similar motion structures. Therefore, motion information of the pixel, say, p_i can be approximated by the neighboring pixels in the same frame. The initial motion vector of the current pixel is approximated by the motion activity of the upper-left neighboring pixels in the same frame.

Prediction Coding

An image normally requires an enormous storage. To transmit an image over a 28.8 Kbps modem would take almost 4 minutes. The purpose for image compression is to reduce the amount of data required for representing images and therefore reduce the cost for storage and transmission. Image compression plays a key role in many important applications, including image database, image communications, remote sensing (the use of satellite imagery for weather and other earth-resource application). The image(s) to be compressed are gray scale with pixel values between 0 to 255. There are different techniques for compressing images. They are broadly classified into two classes called lossless and lossy compression techniques. As the name suggests in lossless compression techniques, no information regarding the image is lost. In other words, the reconstructed image from the compressed image is identical to the original image in every sense. Whereas in lossy compression, some image information is lost, i.e. the reconstructed image from the compressed image is similar to the original image but not identical to it. The temporal prediction residuals from adaptive prediction are encoded using Huffman codes. Huffman codes are used for data compression that will use a variable length code instead of a fixed length code, with fewer bits to store the common characters, and more bits to store the rare characters. The idea is that the frequently occurring symbols are assigned short codes and symbols with less frequency are coded using more bits. The Huffman code can be constructed using a tree. The probability of each intensity level is computed and a column of intensity level with descending probabilities is created. The intensities of this column constitute the levels of Huffman code tree. At each step the two tree nodes having minimal probabilities are connected to form an intermediate node. The probability assigned to this node is the sum of probabilities of the two branches. The procedure is repeated until all branches are used and the probability sum is 1.Each edge in the binary tree, represents either 0 or 1, and each leaf corresponds to the sequence of 0s and 1s traversed to reach a particular code. Since no prefix is shared, all legal codes are at the leaves, and

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

decoding a string means following edges, according to the sequence of 0s and 1s in the string, until a leaf is reached. The code words are constructed by traversing the tree from root to its leaves. At each level 0 is assigned to the top branch and 1 to the bottom branch. This procedure is repeated until all the tree leaves are reached. Each leaf corresponds to a unique intensity level. The codeword for each intensity level consists of 0s and 1s that exist in the path from the root to the specific leaf.

II. TECHNIQUE

The problem laid in the past decades in identifying the unrecognized objects from a high-resolution image. If the image is created from a hyper spectral camera, the problem still laid in identifying what actually the object was, since the hyper spectral image detects only the presence of an object, not what an object actually is. Various derivations [2] and performance [3] computing methods were used in order to obtain the specific property of the image. But since the above methods does not specify what the object property was, there should be a method in order to specify what the object in an image actually was. Since the image taken from a hyper spectral camera suffers from low resolution, we could not identify what actually the particular object was, even though it detects the presence of an object. There is a need for image applications in the detection of objects from a distant location. Normally, the image would be such that the presence of an object could not be detected from it. But, from a hyper spectral camera, the object, if it was on that location, could be captured in the hyper spectral camera. Also, an image taken from a hyper spectral camera suffers from low resolution and thus does not show the exact properties of an image. Since the identification of moving object in a camera is not possible from distant location, to overcome this problem we can use Hyper spectral camera to identify the object. But Hyper spectral camera will only provide the presence of objects, but not what object is. Thus, the problem areas are such that there should be a methodology in identifying an object from a high-resolution image. That is, it should detect the points from a hyper spectral image which are the points that specify the particular objects in the image. The points that resembles the object in the hyper spectral image should be able to be used in retrieving the objects from the high-resolution image. since the objects emits various amounts of energies depending upon the type of objects, they should be identified by showing the presence of it. A variety of simple interpolation methods, such as Pixel Replication, Nearest Neighbour Interpolation, Bilinear Interpolation and Bi-cubic Interpolation have been widely used for CFA demosaicking. But these simple algorithms produce low quality images. More complicated algorithms like the edge-directed interpolation have generated better quality image than simple interpolation methods. But these algorithms still generate the artefacts. Some algorithms have been developed to improve these problems. These algorithms often require huge computation power, so it is impossible to be implemented in real time system. Secondly, images and videos need to be in a compressed form when they have to be send it from source to destination since the image and video data may be huge since it may be containing high resolution data. Thus there is a need for compressing the data thereby reducing its size and thereby making the data efficient to be transferable from source to destination. But the problems arise from the fact that the data when decompressed at the destination should be the same as that of the original data and if it is not obtained as the same, then the compression of the data makes no use. So, the problem lays in providing efficient compression techniques [28][29][34]in order to retrieve the data as same as the original data.

III. DATA

The problem areas are divided into,

1. Target detection and classification of the objects on a specific region.

2. Calculating the frame rates and using compression/decompression techniques to send and retrieve video. .

To handle the problem of Target detection, the Hyper spectral analysis is used. That is, it is used to identify the objects and its background. The background of an object will be always constant. Since the object emits various amounts of energies, the energy analysis of the object is made. If the object is moving then there will be varying amount of emissions for the objects. That will be analysed. Since the background is a constant, and the objects which are moving emits various amounts of energies, the objects can be identified using energy analysis. The precision/accuracy of the object is the case in order to detect the target. For that, the hyper spectral analysis is used in order to identify the background of the object. Smoothening of objects in an image can be done by using filter arrays so that the manipulation of the concerned object by the receiver, when an image is received, can be effectively carried out. The problems related to identifying the object at skylight is handled by the following methods: The first method uses the reflection property of the objects. Since the reflection properties of various objects are different, then it means that various emissions are been made by different objects and by this way, the objects can be identified by these different energy emissions. The second method such as the spectral feature analysis is used to analyze the spectral images. This is used to identify the background from the object since the background is a constant. The third method is mean shift tracking algorithm[22][23][25]. This is used to identify the presence of the object in different frames to know whether the object is moving or not. The fourth method is the tracking algorithm which is used to detect the background and the objects in order to know the presence of objects. The fifth method such as target representation is used to detect the object at a particular target. It uses methods which compares the threshold values to distinguish between background and the object in order to identify it. The threshold value will be set to

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

a value. If the value is less than the threshold, then it will be a background else it will be an object. Lossless JPEG transcoding has many other relevant applications besides reencoding and rotating. For example, it can be used by editing software to avoid a quality loss in the unedited parts of the image. With some additional modifications, it can also be used to perform other simple geometric transformations on JPEG compressed images[34], like cropping or mirroring. Usage of the JPEG file format and the Huffman encoding, nothing else from the JPEG algorithm, therefore the compression scheme is lossless.The transmission of compression images is done using transcoding techniques in order to successively compress and transmitting the data and decompress them in order to obtain the original image.

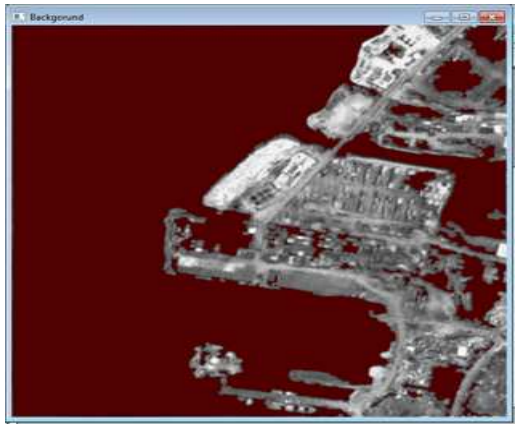


Figure 3. Example of an image with background removed

IV. FIGURES

Object detection

Figure 1. Original image

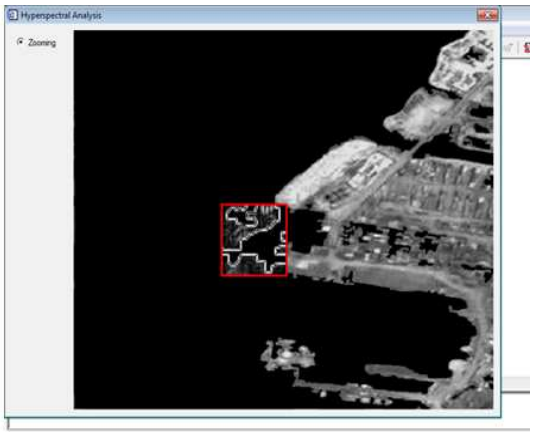


Figure 4. To zoom a particular location in the image

Figure 2. Image converted to grayscale

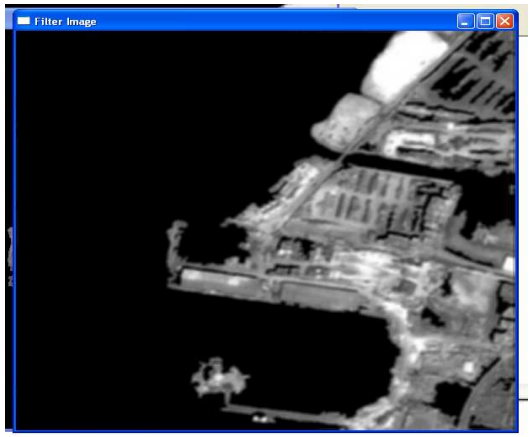


Figure 5. Example of an image smoothened

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Tracking Objects

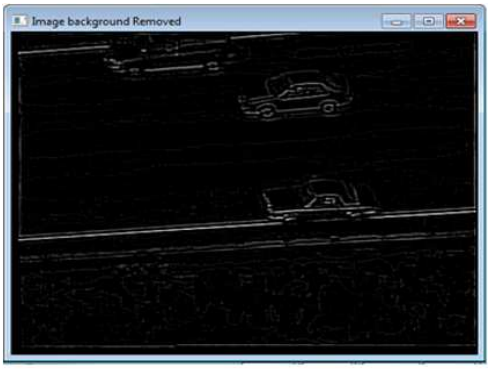


Figure 1. Background removal from frame

Figure 2. Object tracing

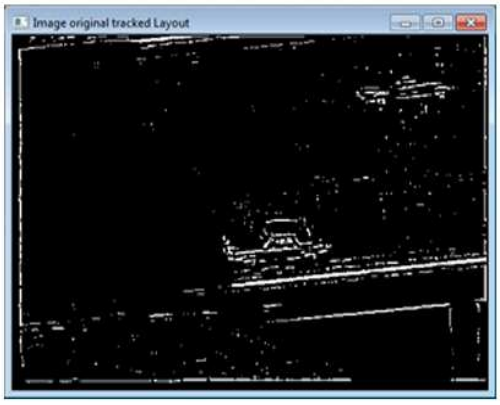


Figure 3. Tracking the moving object

Figure 4. Final result

Figure 5. Tracking of objects in the frame

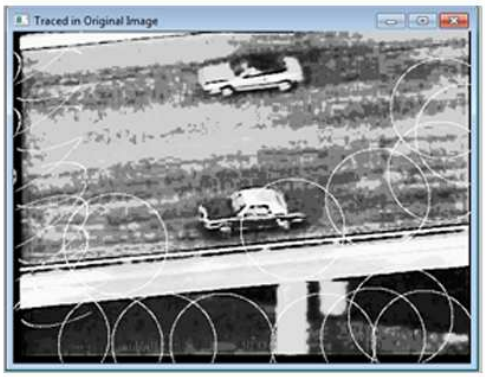


Figure 6. Original Frame used to track object

Figure 7. Replicate image used to track object

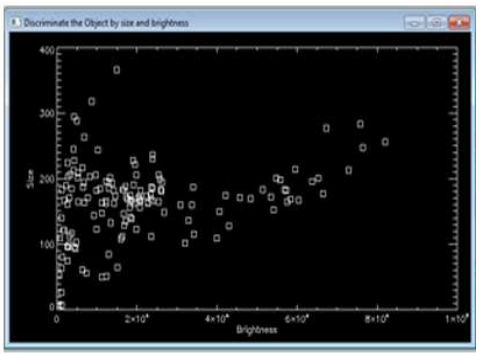
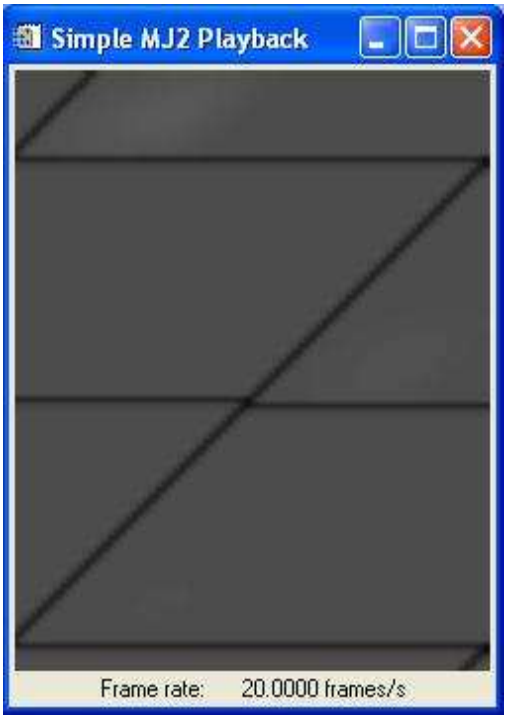


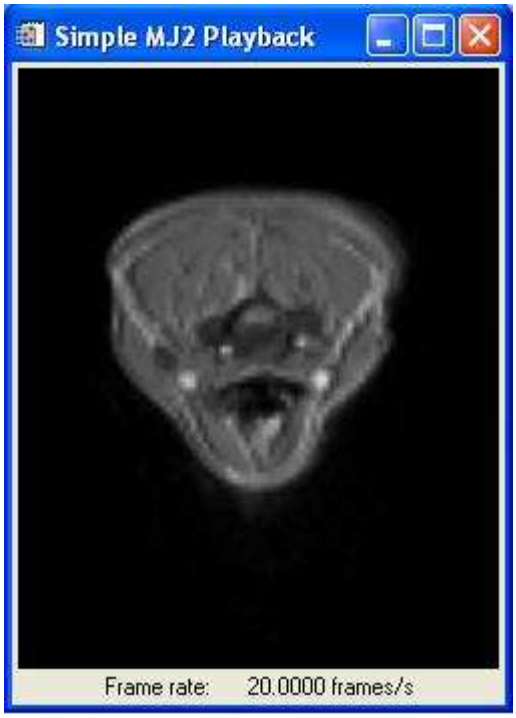
Figure 8. Object discrimination by size and brightness

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

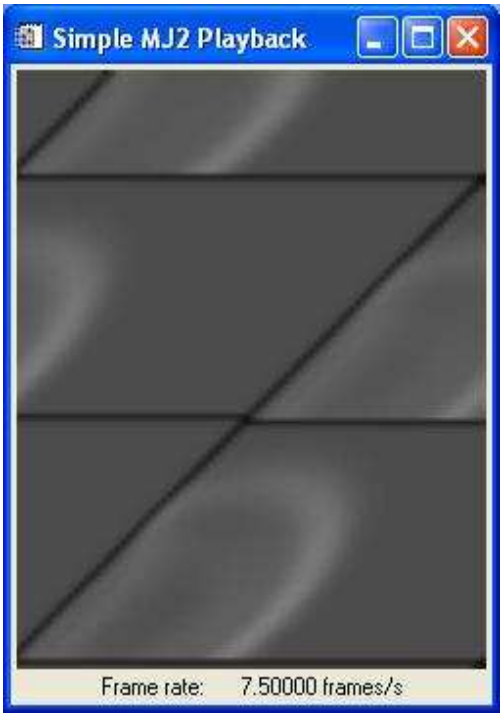
Frame Rate Calculation



Frame rate calculations (original frame rate)



Frame rate calculations (original frame rate)



Frame rate calculations (obtained frame rate)

Frame rate calculations (obtained frame rate)

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

V. CONCLUSIONS

Recent advances in imaging and computer hardware technology have led to an explosion in the use of multispectral, hyper spectral, and in particular, color images/video in a variety of fields including agriculture, ecology, geology, medicine, meteorology, mining, and oceanography. As a result, automated processing and analysis of multichannel images/video have become an active area of research. The volume of data available from both airborne and spaceborne sources will increase rapidly. High resolution hyper spectral remote sensing systems may offer hundreds of bands of data. Efficient use, transmission, storage, and manipulation of such data will require some type of bandwidth compression. Current image compression standards are not specifically optimized to accommodate hyper spectral data. To ensure that the frames when send to the receiver will contain smoother edges for objects, transcoding technique is applied. It uses the concept of replicate array with filter array in order to ensure that the frames are send correctly at the receiver making the object in each frame more identifiable. This ensures that the frames when send from the source will be correctly received at the receiver. The filter array is used because there will be a guarantee that the pixels arrived at the destination will contain adequate information. There is a chance that some of the pixels may be corrupt in the image that is to be send to a destination. So, in order to avoid corrupt pixel values to be send to a destination, the image thus needs to be smoothened out.

ACKNOWLEDGMENT

The authors would like to thank the members of the Dept. of CITE, M S University, Tirunelveli for various algorithms used in this research, and all the people who helped in preparing and carrying out the experiments and data collection.

REFERENCES

[1] Bayesian Estimation of Linear Mixtures Using the Normal Compositional Model. Application to Hyperspectral Imagery Olivier Eches, Nicolas Dobigeon, *Member, IEEE*, Corinne Mailhes, *Member, IEEE*, and Jean-Yves Tourneret, *Senior Member, IEEE*

[2] Image-Derived Prediction of Spectral Image Utility for Target Detection Applicationsb Marcus S. Stefanou, *Member, IEEE*, and John P. Kerekes, *Senior Member, IEEE*

[3]HIGH PERFORMANCE COMPUTING FOR HYPERSPECTRAL IMAGE ANALYSIS: PERSPECTIVE AND STATE-OF-THE-ART *Antonio Plaza*1, *Qian Du*2, *Yang-Lang Chang*3\

[4] Target Detection and Verification via Airborne Hyperspectral and High-Resolution Imagery Processing and Fusion Doron E. Bar, Karni Wolowelsky, Yoram Swirski, Zvi Figov, Ariel Michaeli, Yana Vaynzof, Yoram Abramovitz, Amnon Ben-Dov, Ofer Yaron, Lior Weizman, and Renen Adar

[5] HYPERSPECTRAL TARGET DETECTION FROM INCOHERENT PROJECTIONS Kalyani Krishnamurthy, Maxim Raginsky and Rebecca Willett Department of Electrical and Computer Engineering Duke University, Durham, NC 27708

[6] HYPERSPECTRAL IMAGE ENHANCEMENT WITH VECTOR BILATERAL FILTERING *Honghong Peng* Center for Imaging Science Rochester Institute of Technology 54 Lomb Memorial Drive Rochester NY 14623 *Raghuveer Rao* Army Research Laboratory AMSRD-ARL-SE-SE 2800 Powder Mill Road Adelphi, MD 20783

[7] I. R. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical attern with unknown spectral distribution," *IEEE Trans.Acoust., Speech Signal Process.*, vol. 38, no. 10, pp. 1760–1770, Oct. 1990.

[8] Z. W. Kim and R. Nevatia, "Uncertain reasoning and learning for feature grouping," *Comput. Vis. Image Understanding*, vol. 76, pp. 278–288, 1999.

[9] C. G. Simi, E. M. Winter, M. J. Schlangen, and A. B. Hill, S. S. Shen and M. R. Descour, Eds., "On-board processing for the COMPASS, algorithms for multispectral, hyperspectral, and ultraspectral imagery VII, *Proc. SPIE*, vol.4381, pp.137–142, 2001.

[10] O. Kuybeda, D. Malah, and M. Barzohar, "Rank estimation and redundancy reduction of high-dimensional noisy signals with preservation of rate vectors," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5579–5592, Dec. 2007.

[11] Z. Figov, K.Wolowelsky, and N. Goldberg, L. Bruzzone, Ed., "Co-registration of hyperspectral bands," *Image Signal Process. Remote Sens. XIII. Proc. SPIE*, vol. 6748, pp. 67480s-1–67480s-12, 2007.

[12] L. Boker, S. R. Rotman, and D. G. Blumberg, "Coping with mixtures of backgrounds in a sliding window anomaly detection algorithm," in *Proc. SPIE, Electro-Opt. Infrared Syst.: Technol. Appl. V*, 2008, vol. 7113, pp. 711315-1–711315-12.

[13] C.Cafforio, F.Rocca .Methods for measuring small displacements of television images., *IEEE Trans. on Information Theory*, Vol. IT-22, No. 5, pp. 573-579, Sept. 1976.

[14] R.Moorhead, S.Rajala .Motion-compensated interframe coding., *Proc. ICASSP*, pp. 347-350, 1985.

[15] H.G.Musmann, P.Pirsch, H.-J.Grallert .Advances in Picture Coding., *Proc. of the IEEE*, Vol. 73, No. 4, pp. 523-548, Apr. 1985.

[16] G. Healey and D. Slater. Global color constancy: recognition of objects by use of illumination-invariant properties of color distributions. *JOSAA*, 11(11):3003–3010, 1994.

[17] S. Subramanian and N. Gat. Subpixel object detection using hyperspectral imaging for search and rescue operations. volume 3371, pages 216–225. SPIE, 1998. *SPIE, Electro-Opt. Infrared Syst.: Technol. Appl. V*, 2008, vol. 7113, pp. 711315-1–711315-12.

[18] G. Smith and E.Milton. The use of the empirical line method to calibrate remotely sensed data to reflectance. *Int. J. Remote Sensing,*, 20(13):2653–2662, 1999.

[19] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, pages II: 142–149, 2000.

[20] G. D. Finlayson. Computational color constancy. *International Conference on Pattern Recognition,*, 1:1191, 2000. [12] D. Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *JCSS*, 66(4):671 – 687, 2003. Special Issue on PODS 2001.

[21] M. Lewis, V. Jooste, and A. de Gasparis. Discrimination of arid vegetation with airborne multispectral scanner hyperspectral imagery. *GRS, IEEE Transactions on*, 39(7):1471 –1479, jul 2001.

[22] D. Stein, S. Beaven, L. Hoff, E. Winter, A. Schaum, and A. Stocker. Anomaly detection from hyperspectral imagery. *Sig. Proc. Magazine, IEEE*, 19(1):58 –69, jan 2002.

[23] D. Comaniciu, V. Ramesh, and P.Meer. Kernel-based object tracking. *IEEE Trans. PAMI*, 25(5):564–575, 2003.

[24] D. Manolakis. Detection algorithms for hyperspectral imaging applications: a signal processing perspective. In *Advances in Techniques for Analysis of Remotely Sensed Data, 2003 IEEE Workshop on*, pages 378 – 384, oct. 2003.

[25] C. Shan, Y. Wei, T. Tan, and F. Ojardias. Real time hand tracking by combining particle filtering and mean shift. In *FGR*, pages 669–674. IEEE Computer Society, 2004.

[26] R. Marion, R. Michel, and C. Faye. Measuring trace gases in plumes from hyperspectral remotely sensed data. *GRS, IEEE Transactions on*, 42(4):854 – 864, april 2004.

[27] R.W. R. N. Clark, G. A. Swayze. Sgs digital spectral library splib06a. *U.S. Geological Survey, Data Series 231*, 2007. [23] J. R. Schott. *Remote Sensing: The Image Chain Approach*. Oxford University Press, New York, New York, United States, 2nd edition edition, 2007.

[28] J.G. Apostolopoulos and S.J. Wee, ``Video Compression Standards", *Wiley Encyclopedia of Electrical and Electronics Engineering*, John Wiley & Sons, Inc., New York, 1999.

[29] V. Bhaskaranand K. Konstantinides, *Image and Video Compression Standards: Algorithms and Architectures*, Boston, Massachusetts: KluwerAcademic Publishers, 1997.

[30] J.L. Mitchell, W.B. Pennebaker, C.E. Fogg, and D.J. LeGall, *MPEG Video Compression Standard*, New York: Chapman & Hall, 1997.

[31] B.G. Haskell, A. Puri, A.N. Netravali, *Digital Video: An Introduction to MPEG-2,* KluwerAcademic Publishers, Boston, 1997.

[32] Deng G., Ye H.: Lossless image compression using adaptive predictor combination symbol mapping and context ¯ltering. Proceedings of the IEEE International Conference on Image Processing, Kobe, Japan, Oct. 1999, Vol. 4, pp. 63{7.

[33] Li X., Orchard M. T.: Edge-Directed Prediction for Lossless Compression of Natural Images. IEEE Transactions on Image Processing, June 2001, Vol. 10(6), pp. 813{17.

[34] ITU-T; ISO/IEC: Information technology|JPEG 2000 image coding system: Core coding system. ITU-T Recommendation T.800 and ISO/IEC International Standard 15444-1, August 2002.

[35] Christopoulos C.; Skodras A.; Ebrahimi T.: The JPEG2000 Still Image Coding System an Overview. IEEE Transactions on consumer Electronics, November 2000, Vol. 46(4), pp. 1103{27.

[36] Howard, P. G.; Vitter, J. S.: Fast and e±cient lossless image compression. Pro-ceedings DCC '93 Data Compression Conference, IEEE Comput. Soc. Press, Los Alamitos, California, 1993, pp. 351{60.

[37] Starosolski, R.: Fast, robust and adaptive lossless image compression. Machine Graphics and Vision, 1999, Vol. 8, No. 1, pp. 95-116.

[38] Starosolski, R.; Skarbek, W.: Modi¯ed Golomb{Rice Codes for Lossless Compression of Medical Images. Proceedings of International Conference on E-health in Common Europe, Cracow, Poland, June 2003, pp. 423{37.

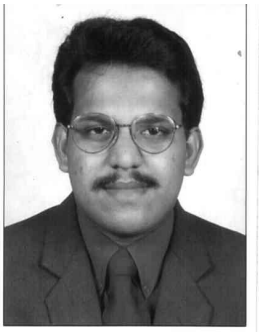
[39] Starosolski, R.: Reversing the Order of Codes in the Rice Family. Studia Informatica, 2002, Vol. 23, No. 4(51), pp. 7{16.

AUTHORS

T. Arumuga Maria Devi received B.E. Degree in Electronic and Communication Engineering from Manonmaniam Sundaranar University, Tirunelveli India in 2003, M.Tech degree in Computer and Information Technology from Manonmaniam Sundaranar University, Tirunelveli, India in 2005. Currently, she is doing Ph.D in Computer and Information Technology and also the Assistant Professor of Centre for Information Technology and Engineering of Manonmaniam Sundaranar University. Her research interests include Signal and Image Processing, Multimedia and Remote Communication.

Nallaperumal Krishnan received M.Sc. degree in Mathematics from Madurai Kamaraj University,Madurai, India in 1985, M.Tech degree in Computer and Information Sciences from Cochin University of Science and Technology, Kochi, India in 1988 and Ph.D. degree in Computer Science & Engineering from Manonmaniam Sundaranar University, Tirunelveli. Currently, he is the Professor and Head of Department of Center for Information Technology and Engineering of Manonmaniam Sundaranar University. His research interests include Signal and Image Processing, Remote Sensing, Visual Perception, and mathematical morphology fuzzy logic and pattern recognition. He has authored three books, edited 18 volumes and published 25 scientific papers in Journals. He is a Senior Member of the IEEE and chair of IEEE Madras Section Signal Processing/Computational Intelligence / Computer Joint Societies Chapter.

Sherin K. K received M.Sc. Software Engineering Degree from Anna University, Chennai India in 2006, Currently he is doing M.Tech degree in Computer and Information Technology (CIT) from Manonmaniam Sundnmaniam Sundaranar University. His research interest include Image Processing.



Mariadas Ronnie C.P received MCA Degree from Bharathiar University, Coimbatore India in 2001, Currently he is doing M.Tech degree in Computer and Information Technology (CIT) from Manonmaniam Sundaranar University. His research interest include Image Processing.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Efficient Retrieval of Unrecognized Objects from Hyper spectral and High Resolution imagery into Jpeg imagery Processing and Fusion

T. Arumuga Maria Devi [1] *Nallaperumal Krishnan*[2] *Mariadas Ronnie C.P* [3]
Assistant Professor, Dept. of CITE *HOD, SMIEEE, Dept. of CITE* *P G Scholar, Dept. of CITE*
Centre for Information Technology and Engineering
Manonmaniam Sundaranar University, Tirunelveli.
Email: deviececit@gmail.com [1]. Email: krishnan@msuniv.ac.in [2]. Email: mariadasronnie@yahoo.co.in [3].
Phone No:9677996337 Phone No: 9443117397 Phone No:8089713017

***Abstract* - This paper deals mainly with the performance study and analysis of image retrieval techniques for retrieving unrecognized objects from an image using Hyper spectral camera and high-resolution image. The main work identified is that efficient retrieval of unrecognized objects in an image will be made possible using spectral analysis and spatial analysis. The methods used above to retrieve unrecognized object from a high-resolution image are found to be more efficient in comparison with the other image retrieval techniques. The detection technique to identify objects in an image is accomplished in two steps: anomaly detection based on the spectral data and the classification phase, which relies on spatial analysis. At the classification step, the detection points are projected on the high-resolution images via registration algorithms. Then each detected point is classified using linear discrimination functions and decision surfaces on spatial features. The two detection steps possess orthogonal information: spectral and spatial. The objects in an image may require that its edges should be smoother in order to make it detect easily by the receiver when it is send from one machine to another. In order to overcome the problems associated with it, Transcoding technique is used by using filter arrays.**

***Keywords*— Anomaly suspect, spectral and spatial analysis, linear discrimination functions, registration algorithms, filter arrays.**

I. INTRODUCTION

The process of recovering unrecognized objects in an image is a trivial task which finds its need in recognizing objects from a distant location. Since there is a need in retrieving unrecognized objects from a high-resolution image, some form of object extraction method from an image is necessary. Remote sensing, for example is often used for detection of predefined targets, such as vehicles, man-made objects, or other specified objects. Since the identification of moving object in a camera is not possible from distant location, to overcome this problem we can use Hyper spectral camera to identify the object. A new technique is thus applied that combines both spectral and spatial analysis for detection and classification of such targets. Fusion of data from two sources, a hyper spectral cube and a high-resolution image, is used as the basis of this technique. Hyper spectral images supply information about the physical properties of an object while suffering from low spatial resolution. There is another problem in a Hyper spectral image, that, it does not identify what an object is, rather, it will detect the presence of an object. In the case of a high resolution image, since the image is such that it does not show the presence of an object, some sort of mechanism is thus needed. That is why, the fusion of the two, the Hyper spectral image and the high-resolution image are used to successfully retrieve the unrecognized object from an image. The use of high-resolution images enables high-fidelity spatial analysis in addition to the spectral analysis. The detection technique to identify objects in an image is accomplished in two steps: anomaly detection based on the spectral data and the classification phase, which relies on spatial analysis. At the classification step, the detection points are projected on the high-resolution images via registration algorithms. Then each detected point is classified using linear discrimination functions and decision surfaces on spatial features. The two detection steps possess orthogonal information: spectral and spatial. At the spectral detection step, we want very high probability of detection, while at the spatial step, we reduce the number of false alarms. The problem thus relies in the area of identifying a specific area in a high-resolution image to know the presence of objects in that area. Each region selected upon the user's interest should be able to detect any presence of objects in that area. Related to the implementation of Trans coding, the work is as follows. The objective of this work is to study the relationship between the operational domains for prediction, according to temporal redundancies between the sequences to be encoded. Based on the motion characteristics of the inter frames, the system will
adaptively select the spatial or wavelet domain for prediction. Also the work is to develop a temporal predictor which exploits the motion information among adjacent frames using extremely low side information. The proposed temporal predictor has to work without the requirement of the transmission of complete motion vector set and hence much overhead would be reduced due to the omission of motion vectors.

Spatial and Wavelet Domain: Comparison

Image compression has become increasingly of interest in both data storage and data transmission from remote acquisition platforms (satellites or airborne) because, after compression, storage space and transmission time are reduced. So, there is a need to compress the data to be transmitted in order to reduce the transmission time and effectively retrieve the data after it has been received by the receiver. The aim is now to determine the operational mode of image sequence

compression according to its motion characteristics. The candidate operational modes are spatial domain and wavelet domain. The wavelet domain is extensively used for compression due to its excellent energy compaction. However, it is pointed out that motion estimation in the wavelet domain might be inefficient due to shift invariant properties of wavelet transform. Hence, it is unwise to predict all kinds of image sequences in the spatial domain alone or in the wavelet domain alone. Hence a method is introduced to determine the prediction mode of an image sequence adaptively according to its temporal redundancies. The amount of temporal redundancy is estimated by the inter frame correlation coefficients of the test image sequence. The inter frame correlation coefficient between frames can be calculated. If the inter frame correlation coefficients are smaller than a predefined threshold, then the sequence is likely to be a high motion image sequence. In this case, motion compensation and coding the temporal prediction residuals in wavelet domain would be inefficient; therefore, it is wise to operate on the sequence in the spatial mode. Those sequences that have larger inter frame correlation coefficients are predicted in direct spatial domain. The frames that have more similarities with very few motion changes are coded using temporal prediction in integer wavelet domain.

Discrete Wavelet Transform (DWT)

Hyper spectral images usually have a similar global structure across components. However, different pixel intensities could exist among nearby spectral components or in the same component due to different absorption properties of the atmosphere or the material surface being imaged. This means that two kinds of correlations may be found in hyper spectral images: intraband correlation among nearby pixels in the same component, and interband correlation among pixels across adjacent components. Interband correlation should be taken into account because it allows a more compact representation of the image by packing the energy into fewer number of bands, enabling a higher compression performance. There are many technologies which could be applied to remove correlation across the spectral dimension, but two of them are the main approaches for hyper spectral images: the KLT and the DWT Discrete Wavelet Transform. (DWT) is the most popular transform for image-based application. They have lower computational complexity, and they provide interesting features such as component and resolution scalability and progressive transmission. A 2-dimensional wavelet transform is applied to the original image in order to decompose it into a series of filtered sub band images. At the top left of the image is a low-pass filtered version of the original and moving to the bottom right, each component contains progressively higher-frequency information that adds the detail of the image. It is clear that the higher-frequency components are relatively sparse, i.e., many of the coefficients in these components are zero or insignificant. The wavelet transform is thus an efficient way of decorrelating or concentrating the important information into a few significant coefficients. The wavelet transform is particularly effective for still image compression and has been adopted as part of the JPEG 2000 standard and for still image texture coding in the MPEG-4 standard.

Motion Estimation Prediction

By Motion estimation, we mean the estimation of the displacement of image structures from one frame to another. Motion estimation from a sequence of images arises in many application areas, principally in scene analysis and image coding. Motion estimation obtains the motion information by finding the motion field between the reference frame and the current frame. It exploits temporal redundancy of an image sequence, and, as a result, the required storage or transmission bandwidth is reduced by a factor of four. Block matching is one of the most popular and time consuming methods of motion estimation. This method compares blocks of each frame with the blocks of its next frame to compute a motion vector for each block; therefore, the next frame can be generated using the current frame and the motion vectors for each block of the frame. Block matching algorithm is one of the simplest motion estimation techniques that compare one block of the current frame with all of the blocks of the next frame to decide where the matching block is located. Considering the number of computations that has to be done for each motion vector, each frame of the image is partitioned into search windows of size H*W pixels. Each search window is then divided into smaller macro blocks of size, say, 8*8 or 16*16 pixels. To calculate the motion vectors, each block of the current frame must be compared to all of the blocks of the next frame with in the search range and the Mean Absolute Difference for each matching block is calculated. The block with the minimum value of the Mean Absolute Difference is the preferred matching block. The location of that block is the motion displacement vector for that block in current frame. The motion activities of the neighboring pixels for a specific frame are different but highly correlated since they usually characterize very similar motion structures. Therefore, motion information of the pixel, say, p_i can be approximated by the neighboring pixels in the same frame. The initial motion vector of the current pixel is approximated by the motion activity of the upper-left neighboring pixels in the same frame.

Prediction Coding

An image normally requires an enormous storage. To transmit an image over a 28.8 Kbps modem would take almost 4 minutes. The purpose for image compression is to reduce the amount of data required for representing images and therefore reduce the cost for storage and transmission. Image compression plays a key role in many important applications, including image database, image communications, remote sensing (the use of satellite imagery for weather and other earth-resource application). The image(s) to be compressed are gray scale with pixel values between 0 to 255. There are different techniques for compressing images. They are broadly

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

classified into two classes called lossless and lossy compression techniques. As the name suggests in lossless compression techniques, no information regarding the image is lost. In other words, the reconstructed image from the compressed image is identical to the original image in every sense. Whereas in lossy compression, some image information is lost, i.e. the reconstructed image from the compressed image is similar to the original image but not identical to it. The temporal prediction residuals from adaptive prediction are encoded using Huffman codes. Huffman codes are used for data compression that will use a variable length code instead of a fixed length code, with fewer bits to store the common characters, and more bits to store the rare characters. The idea is that the frequently occurring symbols are assigned short codes and symbols with less frequency are coded using more bits. The Huffman code can be constructed using a tree. The probability of each intensity level is computed and a column of intensity level with descending probabilities is created. The intensities of this column constitute the levels of Huffman code tree. At each step the two tree nodes having minimal probabilities are connected to form an intermediate node. The probability assigned to this node is the sum of probabilities of the two branches. The procedure is repeated until all branches are used and the probability sum is 1.Each edge in the binary tree, represents either 0 or 1, and each leaf corresponds to the sequence of 0s and 1s traversed to reach a particular code. Since no prefix is shared, all legal codes are at the leaves, and decoding a string means following edges, according to the sequence of 0s and 1s in the string, until a leaf is reached. The code words are constructed by traversing the tree from root to its leaves. At each level 0 is assigned to the top branch and 1 to the bottom branch. This procedure is repeated until all the tree leaves are reached. Each leaf corresponds to a unique intensity level. The codeword for each intensity level consists of 0s and 1s that exist in the path from the root to the specific leaf.

II. TECHNIQUE

The problem laid in the past decades in identifying the unrecognized objects from a high-resolution image. If the image is created from a hyper spectral camera, the problem still laid in identifying what actually the object was, since the hyper spectral image detects only the presence of an object, not what an object actually is. Various derivations [2] and performance [3] computing methods were used in order to obtain the specific property of the image. But since the above methods does not specify what the object property was, there should be a method in order to specify what the object in an image actually was. Since the image taken from a hyper spectral camera suffers from low resolution, we could not identify what actually the particular object was, even though it detects the presence of an object. There is a need for image applications in the detection of objects from a distant location. Normally, the image would be such that the presence of an object could not be detected from it. But, from a hyper spectral camera, the object, if it was on that location, could be captured in the hyper spectral camera. Also, an image taken from a hyper spectral camera suffers from low resolution and thus does not show the exact properties of an image. Since the identification of moving object in a camera is not possible from distant location, to overcome this problem we can use Hyper spectral camera to identify the object. But Hyper spectral camera will only provide the presence of objects, but not what object is. Thus, the problem areas are such that there should be a methodology in identifying an object from a high-resolution image. That is, it should detect the points from a hyper spectral image which are the points that specify the particular objects in the image. Secondly, the points that resembles the object in the hyper spectral image should be able to be used in retrieving the objects from the high-resolution image. A variety of simple interpolation methods, such as Pixel Replication, Nearest Neighbour Interpolation, Bilinear Interpolation and Bi-cubic Interpolation have been widely used for CFA demosaicking. But these simple algorithms produce low quality images. More complicated algorithms like the edge-directed interpolation have generated better quality image than simple interpolation methods. But these algorithms still generate the artefacts. Some algorithms have been developed to improve these problems. These algorithms often require huge computation power, so it is impossible to be implemented in real time system.

III. DATA

The problem areas are divided into,

1. Target detection on a specific region.
2. Classification of the objects based on that region.
3. Transmission of compressed images to a destination. .

To handle the problem of Target detection, the Hyper spectral analysis is used. That is, it is used to identify the objects and its background. The background of an object will be always constant. Since the object emits various amounts of energies, the energy analysis of the object is made. If the object is moving then there will be varying amount of emissions for the objects. That will be analysed. Since the background is a constant, and the objects which are moving emits various amounts of energies, the objects can be identified using energy analysis. The precision/accuracy of the object is the case in order to detect the target. For that, the hyper spectral analysis is used in order to identify the background of the object. Smoothening of objects in an image can be done by using filter arrays so that the manipulation of the concerned object by the receiver, when an image is received, can be effectively carried out. The transmission of compression images is done using trans coding techniques in order to successively compress and transmitting the data and decompress them in order to obtain the original image.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. FIGURES

Figure 1. Original image

Figure 2. Image converted to grayscale

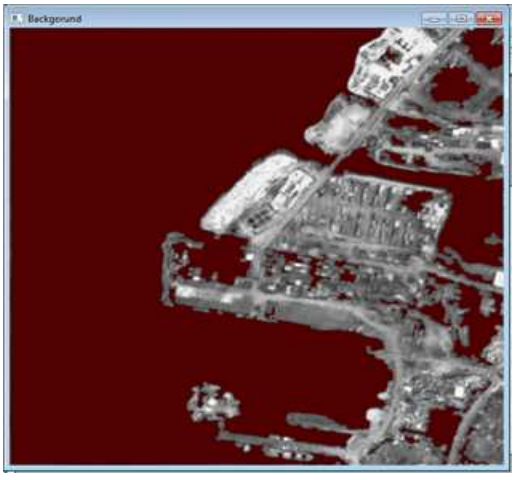


Figure 3. Example of an image with background removal

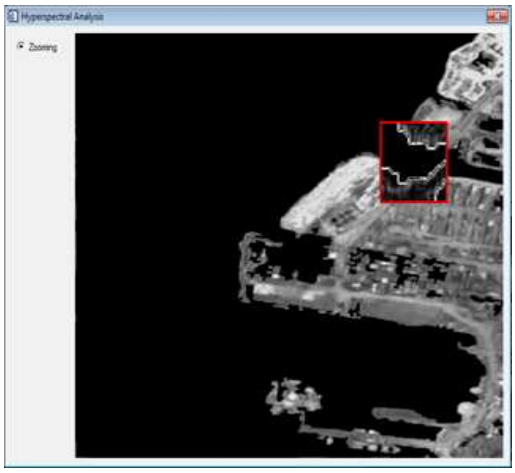


Figure 4. Example of an image that zoomes a location

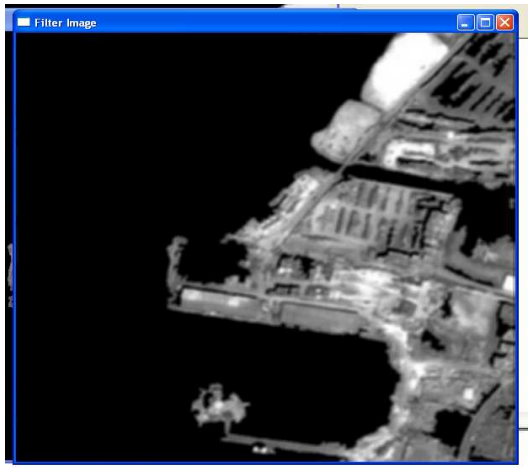


Figure 5. Example of an image smoothened

V. CONCLUSIONS

The classification problem of objects is handled by local detection method to identify the characteristics of the object. Local detection is made by superimposing the points obtained from the hyper spectral image into the high-resolution image there by obtaining the characteristics of the object. Since an accuracy of what object has been identified was not possible on previous methods, a Filter Array is set to identify the background with other objects. These Filter Array will be used to define the pixel information clearly and making these data to be available with less corruption.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

ACKNOWLEDGMENT

The authors would like to thank the members of the Dept. of CITE, M S University, Tirunelveli for various algorithms used in this research, and all the people who helped in preparing and carrying out the experiments and data collection.

REFERENCES

[1] Bayesian Estimation of Linear Mixtures Using the Normal Compositional Model. Application to Hyperspectral ImageryOlivier Eches, Nicolas Dobigeon, *Member, IEEE*, Corinne Mailhes, *Member, IEEE*, and Jean-Yves Tourneret, *Senior Member, IEEE*

[2] Image-Derived Prediction of Spectral Image Utility for Target Detection Applicationsb Marcus S. Stefanou, *Member, IEEE*, and John P. Kerekes, *Senior Member, IEEE*

[3]HIGH PERFORMANCE COMPUTING FOR HYPERSPECTRAL IMAGE ANALYSIS: PERSPECTIVE AND STATE-OF-THE-ART *Antonio Plaza*1, *Qian Du*2, *Yang-Lang Chang*3\

[4] Target Detection and Verification via Airborne Hyperspectral and High-Resolution Imagery Processing and Fusion Doron E. Bar, Karni Wolowelsky, Yoram Swirski, Zvi Figov, Ariel Michaeli, Yana Vaynzof, Yoram Abramovitz, Amnon Ben-Dov, Ofer Yaron, Lior Weizman, and Renen Adar

[5] HYPERSPECTRAL TARGET DETECTION FROM INCOHERENT PROJECTIONS Kalyani Krishnamurthy, Maxim Raginsky and Rebecca Willett Department of Electrical and Computer Engineering Duke University, Durham, NC 27708

[6] HYPERSPECTRAL IMAGE ENHANCEMENT WITH VECTOR BILATERAL FILTERING *Honghong Peng* Center for Imaging Science Rochester Institute of Technology 54 Lomb Memorial Drive Rochester NY 14623 *Raghuveer Rao* Army Research Laboratory AMSRD-ARL-SE-SE 2800 Powder Mill Road Adelphi, MD 20783

[7] I. R. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical attern with unknown spectral distribution," *IEEE Trans.Acoust., Speech Signal Process.*, vol. 38, no. 10, pp. 1760–1770, Oct. 1990.

[8] Z. W. Kim and R. Nevatia, "Uncertain reasoning and learning for feature grouping," *Comput. Vis. Image Understanding*, vol. 76, pp. 278–288, 1999.

[9] C. G. Simi, E. M. Winter, M. J. Schlangen, and A. B. Hill, S. S. Shen and M. R. Descour, Eds., "On-board processing for the COMPASS, algorithms for multispectral, hyperspectral, and ultraspectral imagery VII, *Proc. SPIE*, vol.4381, pp.137–142, 2001.

[10] O. Kuybeda, D. Malah, and M. Barzohar, "Rank estimation and redundancy reduction of high-dimensional noisy signals with preservation of rate vectors," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5579–5592, Dec. 2007.

[11] Z. Figov, K.Wolowelsky, and N. Goldberg, L. Bruzzone, Ed., "Co-registration of hyperspectral bands," *Image Signal Process. Remote Sens. XIII. Proc. SPIE*, vol. 6748, pp. 67480s-1–67480s-12, 2007.

[12] Simple fast and adaptive image compression algorithm.,. *Roman Starosolskil,. Dec 20*, 2006, 2007, 37(1):65-91, DOI: 10.1002/spe.746

[13] Recent trends in image compression and its applications., *IEEE member M.A Ansari and R.S Anand*. XXXII NATIONAL SYSTEMS CONFERENCE, NSC 2008, December 17-19, 2008

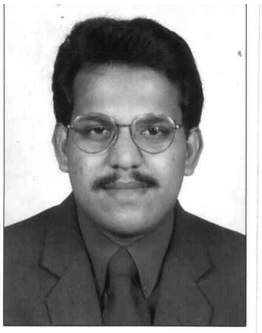
[14] R.Moorhead, S.Rajala .Motion-compensated interframe coding., *Proc. ICASSP*, pp. 347-350, 1985.

[15] The LOCO-I Lossless Image Compression Algorithm: Principles and Standardization into JPEG-LS Marcelo J. Weinberger and Gadiel Seroussi Hewlett-Packard Laboratories, Palo Alto, CA 94304, USA Guillermo Sapiro Department of Electrical and Computer Engineering University of Minnesota, Minneapolis, MN 55455, USA.

AUTHORS

T. Arumuga Maria Devi received B.E. Degree in Electronic and Communication Engineering from Manonmaniam Sundaranar University, Tirunelveli India in 2003, M.Tech degree in Computer and Information Technology from Manonmaniam Sundaranar University, Tirunelveli, India in 2005. Currently, she is doing Ph.D in Computer and Information Technology and also the Assistant Professor of Centre for Information Technology and Engineering of Manonmaniam Sundaranar University. Her research interests include such as Signal and Image Processing and in areas such as Multimedia and Remote Communication.

Nallaperumal Krishnan received M.Sc. degree in Mathematics from Madurai Kamaraj University,Madurai, India in 1985, M.Tech degree in Computer and Information Sciences from Cochin University of Science and Technology, Kochi, India in 1988 and Ph.D. degree in Computer Science & Engineering from Manonmaniam Sundaranar University, Tirunelveli. Currently, he is the Professor and Head of Department of Center for Information Technology and Engineering of Manonmaniam Sundaranar University. His research interests include Signal and Image Processing, Remote Sensing, Visual Perception, and mathematical morphology fuzzy logic and pattern recognition. He has authored three books, edited 18 volumes and published 25 scientific papers in Journals. He is a Senior Member of the IEEE and chair of IEEE Madras Section Signal Processing/Computational Intelligence / Computer Joint Societies Chapter.



Mariadas Ronnie C.P received MCA Degree from Bharathiar University, Coimbatore India in 2001, Currently he is doing M.Tech degree in Computer and Information Technology (CIT) from Manonmaniam Sundaranar University. His research interest include Image Processing.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Retrieving unrecognized objects from HSV into jpeg video at various light resolutions

T. Arumuga Maria Devi [1]
Assistant Professor, Dept. of CITE

Nallaperumal Krishnan[2]
HOD, SMIEEE, Dept. of CITE

K.K Sherin [3]
P G Scholar, Dept. of CITE

Centre for Information Technology and Engineering
Manonmaniam Sundaranar University, Tirunelveli.

Email: deviececit@gmail.com [1].
Phone No:9677996337

Email: krishnan@msuniv.ac.in [2].
Phone No: 9443117397

Email: sherinkk83@yahoo.com [3].
Phone No:9442055500

Abstract — This paper deals mainly with the performance study and analysis of image retrieval techniques for retrieving unrecognized objects from an image using Hyper spectral camera at low light resolution. Since the identification of moving object in a camera is not possible in a low light environment as the object has low reflectance due to lack of lights. Using Hyper spectral data cubes, each object can be identified on the basis of object luminosity. Moving object can be identified by identifying the variation in frame value. The main work identified are that efficient retrieval of unrecognized objects in an image will be made possible using Hyper spectral analysis and various other methods such as Estimation of Reflectance, Feature and mean shift tracker, Traced feature located on image, Band pass filter (Background removal) etc. These methods used above to retrieve unrecognized object from a low light resolution are found to be more efficient in comparison with the other image retrieval techniques.

Keywords— Anomaly suspect, mean shift algorithms, spectral detection, .

I. INTRODUCTION

The process of recovering unrecognized objects from an image in low light is a trivial task which finds its need in recognizing objects from a distant location. Since there is a need in retrieving unrecognized objects from the image, some form of object extraction method from an image is necessary. The application of detecting objects from an image is as follows. Here, we focus on the problem of tracking objects through challenging conditions, such as tracking objects at low light where the presence of the object is difficult to identify. For example, an object which is fastly moving on a plane surface in an abrupt weather condition is normally difficult to identify. A new framework that incorporates emission theory to estimate object reflectance and the mean shift algorithm to simultaneously track the object based on its reflectance spectra is proposed. The combination of spectral detection and motion prediction enables the tracker to be robust against abrupt motions, and facilitate fast convergence of the mean shift tracker. Video images are moving pictures which are sampled at frequent intervals usually, 25 frames per second and stored as sequence of frames. A problem, however, is that digital video data rates are very large, typically in the range of 150 Megabits/second. Data rates of this magnitude would consume a lot of the bandwidth in transmission, storage and computing resources in the typical personal computer. Hence, to overcome these issues, Video Compression standards have been developed and intensive research is going on to derive effective techniques to eliminate picture redundancy, allowing video information to be transmitted and stored in a compact and efficient manner[6].A video image consists of a time-ordered sequence of frames of still images as in figure 1. Generally, two types of image frames are defined: Intra-frames (I-frames) and Inter-frames (P- frames). I-frames are treated as independent key images and P-frames are treated as Predicted frames. An obvious solution to video compression would be predictive coding of P-frames based on previous frames and compression is made by coding the residual error. Temporal redundancy removal is included in P-frame coding, whereas I-frame coding performs only spatial redundancy removal.

II. TECHNIQUE

The problem laid in the past decades in identifying the unrecognized objects from a low light resolution. If the image is created from a hyper spectral camera, the problem still laid in identifying what actually the object was, since the hyper spectral image detects only the presence of an object, not what an object actually is. Various reflectance [24] methods were used in order to obtain the specific property of the image. But since the above methods does not specify what the object property was, there should be a method in order to specify what the object in an image actually was. Since the image taken from a hyper spectral camera suffers from low resolution, we could not identify what actually the particular object was, even though it detects the presence of an object. There is a need for image applications in the detection of objects from a distant location. Normally, the image would be such that the presence of an object could not be detected from it. But, from a hyper spectral camera, the object, if it was on that location, could be captured in the hyper spectral camera. Also, an image taken from a hyper spectral camera suffers from low resolution and thus does not show the exact properties of an image. Since the identification of moving object in a camera is not possible from distant location, to overcome this problem we can use Hyper spectral camera to identify the object.. Thus, the problem areas are such that there should be a methodology in identifying an object from a low light resolution. That is, it should detect the points from a hyper spectral image which are the points that specify the particular objects in the image by reflectance mechanisms of the object. The next problem is such that if an object is fastly moving on a plane surface, it is not necessary that the object will be present on every frame. The points that resembles the object in the hyper spectral image should be able to be used in retrieving the objects by using background removal. Related to the implementation of transcoding, the work is as follows.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The objective of this work is to study the relationship between the operational domains for prediction, according to temporal redundancies between the sequences to be encoded. Based on the motion characteristics of the inter frames, the system will adaptively select the spatial or wavelet domain for prediction. Also the work is to develop a temporal predictor which exploits the motion information among adjacent frames using extremely low side information.
The proposed temporal predictor has to work without the requirement of the transmission of complete motion vector set and hence much overhead would be reduced due to the omission of motion vectors.

Adaptive Domain Selection

This step aims to determine the operational mode of video sequence compression according to its motion characteristics. The candidate operational modes are spatial domain and wavelet domain. The wavelet domain is extensively used for compression due to its excellent energy compaction. However, it is pointed out that motion estimation in the wavelet domain might be inefficient due to shift invariant properties of wavelet transform. Hence, it is unwise to predict all kinds of video sequences in the spatial domain alone or in the wavelet domain alone. Hence a method is introduced to determine the prediction mode of a video sequence adaptively according to its temporal redundancies. The amount of temporal redundancy is estimated by the inter frame correlation coefficients of the test video sequence. The inter frame correlation coefficient between frames can be calculated. If the inter frame correlation coefficients are smaller than a predefined threshold, then the sequence is likely to be a high motion video sequence. In this case, motion compensation and coding the temporal prediction residuals in wavelet domain would be inefficient; therefore, it is wise to operate on the sequence in the spatial mode. Those sequences that have larger inter frame correlation coefficients are predicted in direct spatial domain. The frames that have more similarities with very few motion changes are coded using temporal prediction in integer wavelet domain.

Discrete Wavelet Transform

Discrete Wavelet Transform (DWT) is the most popular transform for image-based application [14], [16], [18]. A 2-dimensional wavelet transform is applied to the original image in order to decompose it into a series of filtered sub band images. At the top left of the image is a low-pass filtered version of the original and moving to the bottom right, each component contains progressively higher-frequency information that adds the detail of the image. It is clear that the higher-frequency components are relatively sparse, i.e., many of the coefficients in these components are zero or insignificant. The wavelet transform is thus an efficient way of decorrelating or concentrating the important information into a few significant coefficients. The wavelet transform is particularly effective for still image compression and has been adopted as part of the JPEG 2000 standard [8] and for still image texture coding in the MPEG-4 standard.

Temporal Residual Prediction

Motion estimation obtains the motion information by finding the motion field between the reference frame and the current frame. It exploits temporal redundancy of video sequence, and, as a result, the required storage or transmission bandwidth is reduced by a factor of four. Block matching is one of the most popular and time consuming methods of motion estimation. This method compares blocks of each frame with the blocks of its next frame to compute a motion vector for each block; therefore, the next frame can be generated using the current frame and the motion vectors for each block of the frame. Block matching algorithm is one of the simplest motion estimation techniques that compare one block of the current frame with all of the blocks of the next frame to decide where the matching block is located. Considering the number of computations that has to be done for each motion vector, each frame of the video is partitioned into search windows of size H*W pixels. Each search window is then divided into smaller macro blocks of size 8*8 or 16*16 pixels. To calculate the motion vectors, each block of the current frame must be compared to all of the blocks of the next frame with in the search range and the Mean Absolute Difference (MAD) for each matching block is calculated. Where N*N is the block size, x(i,j) is the pixel values of current frame at (i,j) th position and y(i+m,j+n) is the pixel value of reference frame at (i+m,j+n) th position. The block with the minimum value of the Mean Absolute Difference (MAD) is the preferred matching block. The location of that block is the motion displacement vector for that block in current frame. The motion activities of the neighboring pixels for aspecific frame are different but highly correlated since they usually characterize very similar motion structures. Therefore, motion information of the pixel $p_i(x,y)$ can be approximated by the neighboring pixels in the same frame. The initial motion vector (Vx, Vy) of the current pixel is approximated by the motion activity of the upper-left neighboring pixels in the same frame.

Coding the Prediction Residual

The temporal prediction residuals from adaptive prediction are encoded using Huffman codes. Huffman codes are used for data compression that will use a variable length code instead of a fixed length code, with fewer bits to store the common characters, and more bits to store the rare characters. The idea is that the frequently occurring symbols are assigned short codes and symbols with less frequency are coded using more bits. The Huffman code can be constructed using a tree. The probability of each intensity level is computed and a column of intensity level with descending probabilities is created. The intensities of this column constitute the levels of Huffman code tree. At each step the two tree nodes having minimal probabilities are connected to form an intermediate

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

node. The probability assigned to this node is the sum of probabilities of the two branches. The procedure is repeated until all branches are used and the probability sum is 1.Each edge in the binary tree, represents either 0 or 1, and each leaf corresponds to the sequence of 0s and 1s traversed to reach a particular code. Since no prefix is shared, all legal codes are at the leaves, and decoding a string means following edges, according to the sequence of 0s and 1s in the string, until a leaf is reached. The code words are constructed by traversing the tree from root to its leaves. At each level 0 is assigned to the top branch and 1 to the bottom branch. This procedure is repeated until all the tree leaves are reached. Each leaf corresponds to a unique intensity level. The codeword for each intensity level consists of 0s and 1s that exist in the path from the root to the specific leaf.

III. DATA

The problem areas are divided as follows:

1. Identifying objects in skylight (during night)
2. To ensure frame clarity

The problems related to identifying the object at skylight is handled by the following methods: The first method uses the reflection property of the objects. Since the reflection properties of various objects are different, then it means that various emissions are been made by different objects and by this way, the objects can be identified by these different energy emissions. The second method such as the spectral feature analysis is used to analyze the spectral images. This is used to identify the background from the object since the background is a constant. The third method is mean shift tracking algorithm. This is used to identify the presence of the object in different frames to know whether the object is moving or not. The fourth method is the tracking algorithm which is used to detect the background and the objects in order to know the presence of objects. The fifth method such as target representation is used to detect the object at a particular target. It uses methods which compares the threshold values to distinguish between background and the object in order to identify it. The threshold value will be set to a value. If the value is less than the threshold, then it will be a background else it will be an object.

Lossless JPEG transcoding has many other relevant applications besides reencoding and rotating. For example, it can be used by editing software to avoid a quality loss in the unedited parts of the image. With some additional modifications, it can also be used to perform other simple geometric transformations on JPEG compressed images, like cropping or mirroring. Usage of the JPEG file format and the Huffman encoding, nothing else from the JPEG algorithm, therefore the compression scheme is lossless.

IV. FIGURES

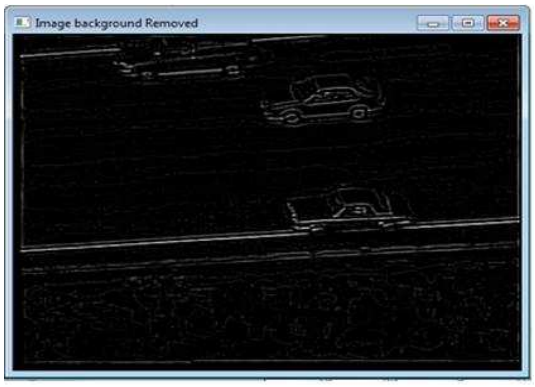


Figure 1. Background removed from a frame

Figure 2. Background removed from another frame

Figure 3. Object tracing

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

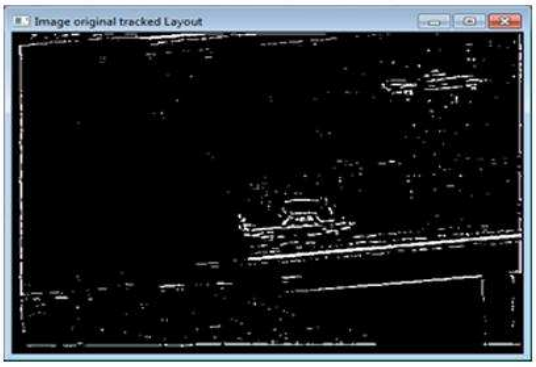


Figure 4. Tracking the moving object

Figure 5. Final result

Figure 6. Tracking of objects in the frame

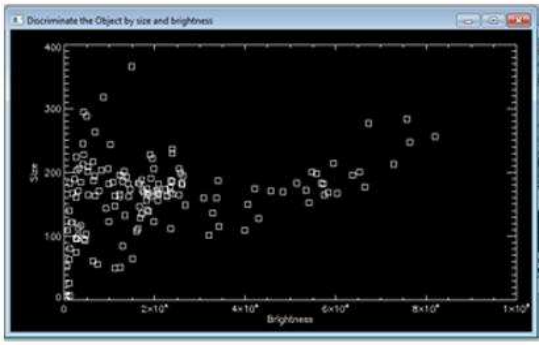


Figure 7. Object discrimination by size and brightness

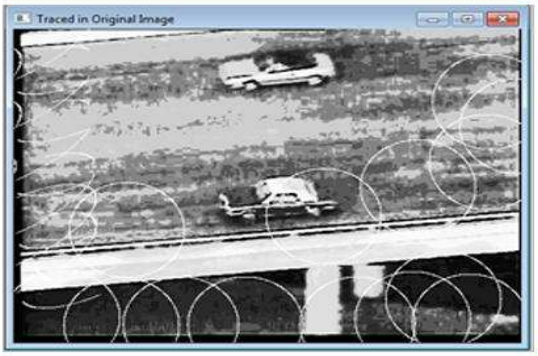


Figure 8. Original Frame used to track object

Figure 9. Replicate image used to track object

V. CONCLUSIONS

The classification problem of objects is handled by local detection method to identify the characteristics of the object. Local detection is made by superimposing the points obtained from the hyper spectral image into the high-resolution image there by obtaining the characteristics of the object. Since an accuracy of what object has been identified was not possible on previous methods, a threshold value is set to identify the background with other objects. The image is first converted from RGB to Gray Scale. Then the pixel values of the image are compared with a threshold value. If the pixel value of the image is below the threshold value, then it is set as a background and is set to 0, else the pixel value is taken as the pixel value for an object and is set to 1. Thus we get an image with unnecessary objects removed by setting it as background and the presence of the object in the image is only shown ensuring frame clarity. To ensure that the frames when send to the receiver will contain smother edges for objects, trans coding technique is applied. It uses the concept of replicate array with filter array in order to ensure that the frames are send correctly at the receiver making the object in each frame more identifiable. This ensures that the frames when send from the source will be correctly received at the receiver.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

ACKNOWLEDGMENT

The authors would like to thank the members of the Dept. of CITE, M S University, Tirunelveli for various algorithms used in this research, and all the people who helped in preparing and carrying out the experiments and data collection.

REFERENCES

[1] Tracking via Object Reflectance using Hyper spectral video camera:
Hein Van Nguyen, Amit Banerjee, Rama Chellappa, Centre for Automation Research, University of Maryland, College Park, Applied Physics Laboratory, Johns Hopkins University.

[2] Transform Coding techniques for lossy hyper spectral data compression
Barbara Penna, Tammam Tillo, Enrico Magli, Gabriela Olmo, Members IEEE

[3] Simple Fast and Adaptive Lossless Image Compression Algorithm
Roman Starosolski, 2007, 37(1):65-91, DOI: 10.1002/spe.746

[4] The LOCO-I Lossless Image Compression Algorithm: Principles and Standardization into JPEG-LS Marcelo J. Weinberger and Gadiel Seroussi., Hewlett-Packard Laboratories, Palo Alto, CA 4304, USA., Guillermo Sapiro., Department of Electrical and Computer Engineering University of Minnesota, Minneapolis, MN 55455, USA.

[5] Recent trends in image compression and its applications
M.A Ansari and R.S Anand. ., IEEE member., XXXII NATIONAL SYSTEMS CONFERENCE, NSC 2008, December 17-19, 2008

[6] Hyper spectral image compression with optimization for spectral analysis
Kameron Romines, Edwin Hong, Assistant Professor, University of Washington

[7] An efficient compression algorithm for hyper spectral images based on a modified coding framework of H.264/AVC
Guizhong Liu, Fan Zhao, Guofu Qu

[8] T. S. J. L. R.A. Neville, K. Staenz and P. Hauff. Automatic endmember extraction from hyperspectral data for mineral exploration. 4th Int. Airborne Remote Sensing, Ottawa, Ontario, Canada, pages 891–896, 1999.

[9] G. Smith and E.Milton. The use of the empirical line method to calibrate remotely sensed data to reflectance. Int. J. Remote Sensing,, 20(13):2653–2662, 1999.

[10] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In CVPR, pages II: 142–149, 2000.

[11] M. Lewis, V. Jooste, and A. de Gasparis. Discrimination of arid vegetation with airborne multispectral scanner hyperspectral imagery. GRS, IEEE Transactions on, 39(7):1471 –1479, jul 2001.

[12] D. Stein, S. Beaven, L. Hoff, E. Winter, A. Schaum, and A. Stocker. Anomaly detection from hyperspectral imagery. Sig. Proc. Magazine, IEEE, 19(1):58 –69, jan 2002.

[13] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. IEEE Trans. PAMI, 24(5):603–619, 2002.

[14] D. Manolakis. Detection algorithms for hyperspectral imaging applications: a signal processing perspective. In Advances in Techniques for Analysis of Remotely Sensed Data, 2003 IEEE Workshop on, pages 378 – 384, oct. 2003.

[15] M. Gianinetto and G. Lechi. The development of superspectral approaches for the improvement of land cover classification. GRS, IEEE Transactions on, 42(11):2670 – 2679, nov. 2004.

[16] C. Shan, Y. Wei, T. Tan, and F. Ojardias. Real time hand tracking by combining particle filtering and mean shift. In FGR, pages 669–674. IEEE Computer Society, 2004.

[17] R. Marion, R. Michel, and C. Faye. Measuring trace gases in plumes from hyperspectral remotely sensed data. GRS, IEEE Transactions on, 42(4):854 – 864, april 2004.

[18] R. Baraniuk, M. Davenport, R. Devore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. Constr. Approx, 2008, 2007.

AUTHORS

T. Arumuga Maria Devi received B.E. Degree in Electronic and Communication Engineering from Manonmaniam Sundaranar University, Tirunelveli India in 2003, M.Tech degree in Computer and Information Technology from Manonmaniam Sundaranar University, Tirunelveli, India in 2005. Currently, she is doing Ph.D in Computer and Information Technology and also the Assistant Professor of Centre for Information Technology and Engineering of Manonmaniam Sundaranar University. Her research interests include Signal and Image Processing, Multimedia and Remote Communication.

Nallaperumal Krishnan received M.Sc. degree in Mathematics from Madurai Kamaraj University,Madurai, India in 1985, M.Tech degree in Computer and Information Sciences from Cochin University of Science and Technology, Kochi, India in 1988 and Ph.D. degree in Computer Science & Engineering from Manonmaniam Sundaranar University, Tirunelveli. Currently, he is the Professor and Head of Department of Center for Information Technology and Engineering of Manonmaniam Sundaranar University. His research interests include Signal and Image Processing, Remote Sensing, Visual Perception, and mathematical morphology fuzzy logic and pattern recognition. He has authored three books, edited 18 volumes and published 25 scientific papers in Journals. He is a Senior Member of the IEEE and chair of IEEE Madras Section Signal Processing/Computational Intelligence / Computer Joint Societies Chapter.

Sherin K. K received M.Sc. Software Engineering Degree from Anna University, Chennai India in 2006, Currently he is doing M.Tech degree in Computer and Information Technology (CIT) from Manonmaniam Sundnmaniam Sundaranar University. His research interest include Image Processing.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Bayesian Spam Filtering using Statistical Data Compression

V. Sudhakar (M.Tech (IT)) .
Avanthi Institute of Engineering and Technology, Visakhapatnam, vsudhakarmtech@yahoo.com

Dr.CPVNJ.Mohan Rao
Professor in CSE Dept, Avanthi Institute of Engineering andTechnology, Visakhapatnam
mohanrao_p@yahoo.com

Satya Pavan Kumar Somayajula
Asst. Professor, CSE Dept, Avanthi Institute of Engineering and Technology, Visakhapatnam,
balasriram1982@gmail.com

Abstract

The Spam e-mail has become a major problem for companies and private users. This paper associated with spam and some different approaches attempting to deal with it. The most appealing methods are those that are easy to maintain and prove to have a satisfactory performance. Statistical classifiers are such a group of methods as their ability to filter spam is based upon the previous knowledge gathered through collected and classified e-mails. A learning algorithm which uses the Naive Bayesian classifier has shown promising results in separating spam from legitimate mail.

Introduction

Spam has become a serious problem because in the short term it is usually economically beneficial to the sender. The low cost of e-mail as a communication medium virtually guaranties profits. Even if a very small percentage of people respond to the spam advertising message by buying the product, this can be worth the money and the time spent for sending bulk e-mails. Commercial spammers are often represented by people or companies that have no reputation to lose. Because of technological obstacles with e-mail infrastructure, it is difficult and time-consuming to trace the individual or the group responsible for sending spam. Spammers make it even more difficult by hiding or forging the origin of their messages. Even if they are traced, the decentralized architecture of the Internet with no central authority makes it hard to take legal actions against spammers. The statistical filtering (especially Bayesian filtering) has long been a popular anti-spam approach, but spam continues to be a serious problem to the Internet society. Recent spam attacks expose strong challenges to the statistical filters, which highlights the need for a new anti-spam approach. The economics of spam dictates that the spammer has to target several recipients with identical or similar e-mail messages. This makes collaborative spam filtering a natural defense paradigm, wherein a set of e-mail clients share their knowledge about recently receivedspame-mails, providing a highly effective defense against a substantial fraction of spam attacks. Also, knowledge sharing can significantly alleviate the burdens of frequent training stand-alone spam filters. However, any large-scale collaborative anti-spam approach is faced with a fundamental and important challenge, namely ensuring the privacy of the e-mails among untrusted e-mail entities. Different from the e-mail service providers such as Gmail or Yahoo mail, which utilizes spam or ham(non-spam) classifications from all its users to classify new messages, privacy is a major concern for cross-enterprise collaboration, especially in a large scale. The idea of collaboration implies that the participating users and e-mail servers have to share and exchange information about the e-mails (including the classification result). However, e-mails are generally considered as private communication between the senders and the recipients, and they often contain personal and confidential information. Therefore, users and organizations are not comfortable sharing information about their e-mails until and unless they are assured that no one else (human or machine) would become aware of the actual contents of their e-mails. This genuine concern for privacy has deterred users and organizations from participating in any large-scale collaborative spam filtering effort. To protect e-mail privacy, digest approach has been proposed in the collaborative anti-spam systems to both provide encryption for the e-mail messages and obtain useful information (fingerprint) from spam e-mail. Ideally, the digest calculation has to be a one-way function such that it should be computationally hard to generate the corresponding e-mail message. It should embody the textual features of the e-mail message such that if two e-mails have similar syntactic structure, then their fingerprints should also be similar.Afew distributed spam identification schemes, such as Distributed Checksum Clearinghouse (DCC) [2] and Vipul's Razor [3] have different ways to generate fingerprints. However, these systems are not sufficient to handle two security threats: 1) Privacy breach as discussed in detail in Section 2 and 2) Camouflage attacks, such as character replacement and good word appendant, make it hard to generate the same e-mail fingerprints for highly similar spam e-mails.

Statistical Data Compression

Probability plays a central role in data compression: Knowing the exact probability distribution governing an information source allows us to construct optimal or near-optimal codes for messages produced by the source. A statistical data compression algorithm exploits this relationship by building a statistical model of the information source, which can be used to estimate the probability of each possible message. This model is coupled with an encoder that uses these probability estimates to construct the final binary representation. For our purposes, the encoding problem is irrelevant. We therefore focus on the source modeling task.

Preliminaries

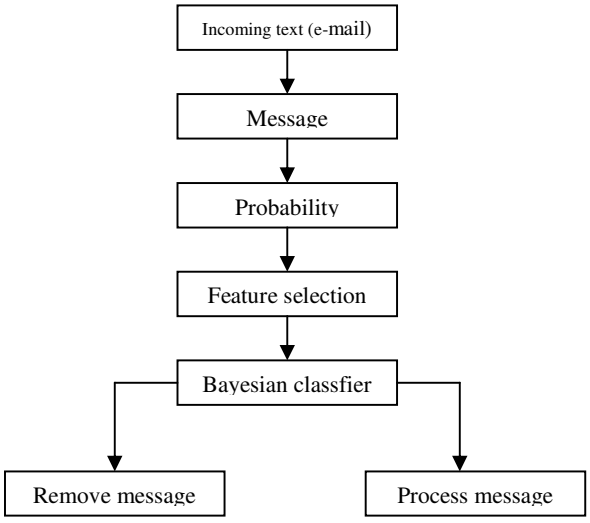
We denote by X the random variable associated with the source, which may take the value of any message the source is capable of producing, and by P the probability distribution

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

over the values of X with the corresponding probability mass function p. We are particularly interested in modeling of text generating sources. Each message **x** produced by such a source is naturally represented as a sequence $\mathbf{X}=x1^n$ = x1….xn $\in \sum^*$of symbols over the source alphabet $\sum$. The length $|x|$ of a sequence can be arbitrary. For text generating sources, it is common to interpret a symbol as a single character, but other schemes are possible, such as binary (bitwise) or word-level models. The entropy $H(X)$ of a source X gives a lower bound on the average per-symbol code length required to encode a message without loss of information: $H(x)=E_{x\sim p}(-\frac{1}{|x|}\log p(x))$ This bound is achievable *only* when the true probability distribution P governing the source is known. In this case, an average message could be encoded using no less than $H(X)$ bits per symbol. However, the true distribution over all possible messages is typically unknown. The goal of any statistical data compression algorithm is then to infer a probability mass function over sequences $f:\sum^*\rightarrow [0,1]$, which matches the true distribution of the source as accurately as possible. Ideally2, a sequence **x** is then encoded with $L(\mathbf{x})$ bits, where $L(\mathbf{x})$ = - log f (**x**). The compression algorithm must therefore *learn* an approximation of P in order to encode messages efficiently. A better approximation will, on average, lead to shorter code lengths. This simple observation alone gives compelling motivation for the use of compression algorithms in text categorization.

Bayesian spam filtering

Bayesian spam filtering can be conceptualized into the model presented in Figure 1. It consists of four major modules, each responsible for four different processes: message tokenization, probability estimation, feature selection and Naive Bayesian classification.



When a message arrives, it is firstly tokenized into a set of features (tokens), F . Every feature is assigned an estimated probability that indicates its spaminess. To reduce the dimensionality of the feature vector, a feature selection algorithm is applied to output a subset of the features. The Naive Bayesian classifier combines the probabilities of every feature in 1 F , and estimates the probability of the message being spam. In the following text, the process of Naive Bayesian classification is described, followed by details concerning the measuring performance. This order of explanation is necessary because the sections concerned with the first three modules require understanding of the classification process and the parameters used to evaluate its improvement.

Performance evolution

Precision and recall a well employed metric for performance measurement in information retrieval is precision and recall. These measures have been diligently used in the context of spam classification (Sahami et al.1998). Recall is the proportion of relevant items that are retrieved, which in this case is the proportion of spam messages that are actually recognized. For example if 9 out of 10 spam messages are correctly identified as spam, the recall rate is 0.9. Precision is defined as the proportion of items retrieved that are relevant. In the spam classification context, precision is the proportion of the spam messages classified as spam over the total number of messages classified as spam. Thus if only spam messages are classified as spam then the precision is 1. As soon as a good legitimate message is classified as spam, the precision will drop below 1. Formally: Let gg n be the number of good messages classified as good (also known as false negatives). Let gs n be the number of good messages classified as spam (also known as false positives).(9). Let ss n be the number of spam messages classified as spam (also known as true positives). Let sg n be the number of spam messages classified as good (also known as true negatives). The precision calculates the occurrence of false positives which are good messages classified as spam. When this happens p drops below 1. Such misclassification could be a disaster for the user whereas the only impact of a low recall rate is to receive spam messages in the inbox. Hence it is more important for the precision to be at a high level than the recall rate. The precision and recall reveal little unless used together. Commercial spam filters sometimes claim that they have an incredibly high precision value of 0.9999% without mentioning the related recall rate. This can appear to be very good to the untrained eye. A reasonably good spam classifier should have precision very close to 1 and a recall rate > 0.8. A problem when evaluating classifiers is to find a good balance between the precision and recall rates. Therefore it is necessary to use a strategy to obtain a combined score. One way to achieve this is to use weighted accuracy.

Cross validation

There are several means of estimating how well the classifier works after training. The easiest and most straightforward means is by splitting the corpus into two parts and using one part for training and the other for testing. This is called the holdout method. The disadvantage is that the evaluation depends heavily on which samples end up in which set. Another method that reduces the variance of the holdout method is k -fold cross-validation. In k -fold cross-validation (Kohavi 1995) the corpus, M , is split into k mutually exclusive parts, M_1,M_2,…….M_k. The inducer is trained on M/M_1 and tested against M_1 . This is repeated k times with different i such that i ∈ {1,2,…k}.Finally the performance is estimated as the mean of the total number of tests.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Conclusion

Optimal search algorithm called SFFS was applied to find a subset of delimiters for the tokenizer. Then a filter and a wrapper algorithm were proposed to determine how beneficial a group of delimiters is to the classification task. The filter approach ran about ten times faster than the wrapper, but did not produce significantly better subsets than the base-lines. The wrapper did improve the performance on all corpuses by finding small subsets of delimiters. This suggested an idea concerning how to select delimiters for a near-optimal solution, namely to start with space and then add a few more. Since the wrapper generated subsets had nothing in common apart from space, the recommendation is to only use space as a delimiter. The wrapper was far too slow to use in spam filter.

References

[1]. Almuallim, H. and T. Dietterich. (1991), Learning with many irrelevant features. In Proceedings of the Ninth National Conference on Artificial Intelligence, pp. 547-552. Menlo Park, CA: AAAI Press/The MIT Press.

[2]. Androutsopoulos I., Paliouras G., Karkaletsis V., Sakkis G., Spyropoulos C. and Stamatopoulos, P. (2000a) Learning to filter spam email: A comparison of a naive bayesian and a memory-based approach. In Workshop on Machine Learning and Textual Information Access, 4th European

[3]. Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000). Androutsopoulos, I., Koutsias, J., Chandrinos, K.V., Paliouras, George and Spyropoulos, C.D. (2000b),

[4]. An Evaluation of Naive Bayesian Anti-Spam Filtering. In Potamias, G., Moustakis, V. and van Someren, M. (Eds.), Proceedings of the Workshop on Machine Learning in the New

[5]. Information Age, 11th European Conference on Machine Learning (ECML 2000), Barcelona, Spain, pp. 9-17.

[6]. Androutsopoulos, I., Paliouras, G., Michelakis, E. (2004), Learning to Filter Unsolicited Commercial E-Mail. Athens University of Economics and Business and National Centre for Scientific Research "Demokritos" Bevilacqua-Linn M. (2003),

[7]. Machine Learning for Naive Bayesian Spam Filter Tokenization Breiman, L., and Spector, P. (1992), Submodel selection and evaluation in regression: The Xrandom case. International Statistical Review, 60, 291-319.

[8]. Androutsopoulos, G. Paliouras, and E. Michelakis. Learning to filter unsolicited commercial e-mail. Technical Report 2004/2, NCSR "Demokritos", October 2004.

[9]. F. Assis, W. Yerazunis, C. Siefkes, and S. Chhabra. CRM114 versus Mr. X: CRM114 notes for the TREC 2005 spam track. In *Proc. 14th Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, November 2005.

[10]. A. R. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998. D. Benedetto, E. Caglioti, and Loreto V. Language trees and zipping. *Physical Review Letters*, 88 (4), 2002.

[11]. A. Bratko and B. Filipi˘c. Spam filtering using character-level markov models: Experiments for the TREC 2005 Spam Track

V.Sudhakar, Studying M.Tech in Information Technology. in CSE Department, Avanthi Institute of Engg & Tech, Tamaram, Visakhapatnam,A.P., India.

Mr. Satya P Kumar Somayajula is working as an Asst.Professor, in CSE Department, Avanthi Institue of Engg & Tech, Tamaram, Visakhapatnam,A.P., India. He has received his M.Sc(Physics) from Andhra University, Visakhapatnam and M.Tech (CST) from Gandhi Institute of Technology And Management University (GITAM University), Visakhapatnam, A.P., INDIA. He published 7 papers in reputed International journals & 5 National journals. His research interests include Image Processing, Networks security, Web security, Information security, Data Mining and Software Engineering.

Dr. C.P.V.N.J. Mohan Rao is Professor in the Department of Computer Science and Engineering and Principal of Avanthi Institute of Engineering & Technology - Narsipatnam. He did his PhD from Andhra University and his research interests include Image Processing, Networks, Information security, Data Mining and Software Engineering. He has guided more than 50 M.Tech Projects and currently guiding four research scholars for Ph.D. He received many honors and he has been the member for many expert committees, member of many professional bodies and Resource person for various organizations.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Mrs Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Mr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Mr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Mr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Mr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Mr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Mr. P. Vasant, University Technology Petronas, Malaysia
Mr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Mr. Praveen Ranjan Srivastava, BITS PILANI, India
Mr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Mr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Mr. Tirthankar Gayen, IIT Kharagpur, India
Ms. Huei-Ru Tseng, National Chiao Tung University, Taiwan

CALL FOR PAPERS
International Journal of Computer Science and Information Security
January - December
IJCSIS 2012
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500